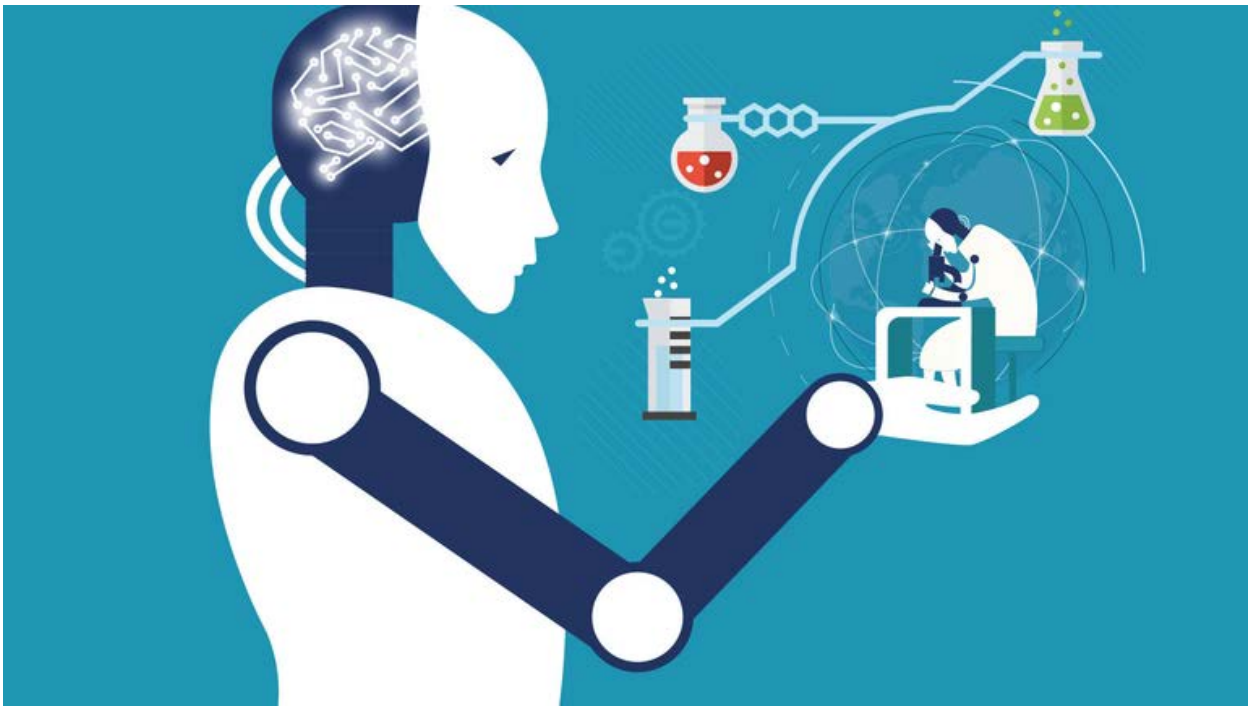
تحلیل و ارزیابی انتقادی «هوش مصنوعی» در پزشکی و مراقبت سلامت از منظر علوم انسانی پزشکی

Critical analysis/evaluation of AI in healthcare and medicine from the perspective of Medical Humanities



منابع حیطهٔ مطالعات میان رشته ای علوم انسانی و سلامت

سیزدهمین المپیاد علمی دانشجویان علوم پزشکی کشور

اسفند ۱۳۹۹



راهنمای مطالعه

دانشجویان عزیز

بسیار خرسندیم که شما عزیزان به حیطه علوم انسانی سلامت علاقه‌مند شده‌اید و المپیاد علمی را زمینه‌ای برای تأمل جدی در آن یافته‌اید. مایه مسرت و خوشوقتی است که یک‌بار دیگر این فرصت فراهم آمده است تا در قالب المپیاد علمی دانشجویان در کنار یکدیگر باشیم. مسلماً برگزاری چنین حرکت علمی عظیمی بدون حضور استادان به‌نام و صاحب‌نظر در کشور و مشارکت فعالانه و خلاقانهٔ شما دانشجویان عزیز میسر نخواهد بود. کمیتهٔ علمی حیطهٔ مطالعات میان‌رشته‌ای علوم انسانی و سلامت همواره بر آن بوده است تا با رعایت بالاترین معیارهای علمی و آکادمیک و بهره‌گیری از به‌روزترین منابع، هم در آموزش و هم در طراحی و داوری آزمون‌ها، انصاف و عدالت را لحاظ کند. کمیتهٔ علمی این حیطه امسال متشکل از جمعی از استادان به‌نام در حوزه‌های هوش‌مصنوعی، فلسفهٔ تحلیلی، فلسفهٔ علم، فلسفهٔ پزشکی، اخلاق پزشکی و آموزش پزشکی هستند.

بی‌تردید، فرآیند المپیاد از حیث آموزشی و پژوهشی ممتاز و متمایز از نظام رسمی آموزش پزشکی است و ملاحظات مختص به خود دارد. ازاین‌رو، فرآیند تعیین منابع و هم‌چنین روند طرح سؤال و نیز سنجش و برگزاری آزمون، مستلزم طرح و بحث مکرر در کمیته علمی است. به دیگر سخن، فرآیند تعیین و گردآوری منابع نیز محتاج تأمل و گفتگوی مکرر اعضای کمیته علمی ست. آنچه در ذیل می‌آید از این مسیر حاصل شده است. لطفاً با دقت به آن توجه بفرمایید و برای موفقیت در سیر المپیاد همه جوانب آن را لحاظ کنید.

مقدمه

هوش‌مصنوعی به‌سرعت در حال گسترش در پزشکی و مراقبت سلامت است. تکنولوژی‌های نوین مبتنی بر هوش‌مصنوعی در تشخیص، درمان و سایر حوزه‌های مراقبت این امکان را به وجود آورده‌اند که مرزهای مرگ و زندگی، بهنجاری و نابهنجاری و سلامت و بیماری به چالش جدی کشیده شوند و پیامدهای خطیر فرهنگی، اجتماعی، اخلاقی، حقوقی در پی داشته باشند. از همین رو رصد، تأمل و نقد جدی و مستمر نقش و جایگاه و پیامدهای تکنولوژی‌های هوش‌مصنوعی در حوزهٔ سلامت بخشی کلیدی از مطالعات میان‌رشته‌ای علوم انسانی و سلامت است. باید توجه کرد هدف در اینجا ارزش‌داوری دربارهٔ تکنولوژی نیست (داوری کنیم که تکنولوژی خوب است یا بد؟)، بلکه بررسی همه جانبهٔ این پدیده و پیامدها و اقتضائات وجودی، فرهنگی، اجتماعی، اخلاقی و... است. از همین رو کمیتهٔ علمی حیطهٔ مطالعات میان‌رشته‌ای علوم انسانی و سلامت بر آن شد تا تحلیل و ارزیابی انتقادی نقش و جایگاه و پیامدهای تکنولوژی هوش‌مصنوعی در حوزهٔ سلامت از منظر علوم انسانی پزشکی را مورد توجه جدی قرار دهد و به‌عنوان موضوع امسال برگزیند.

هوش‌مصنوعی نه از منظر مهندسی، بلکه از منظر علوم انسانی پزشکی

همان‌طور که در عنوان هم آمده است هدف در این المپیاد تحلیل/ ارزیابی انتقادی هوش‌مصنوعی در پزشکی و مراقبت سلامت از منظر علوم انسانی پزشکی است. به‌عنوان‌مثال برای آشنایی بیشتر نگاه کنید به مقاله سروش، الهه، منجمی، علیرضا. (۱۳۹۶). تحلیل و نقد هوش‌مصنوعی در طبابت از منظر معرفت‌شناسی پزشکی. فلسفه علم، ۷(۱۴), ۲۷-۵۸. بدیهی است تحلیل و ارزیابی هوش‌مصنوعی در حوزهٔ سلامت، **از منظر مهندسی نیست**، بلکه جنس تحلیل و ارزیابی از دیدگاه مطالعات میان‌رشته‌ای علوم انسانی و سلامت (medical humanities) است. البته نقد و تحلیل هوش‌مصنوعی از منظر علوم انسانی سلامت نیازمند آن است که با هوش‌مصنوعی در حوزه سلامت و پزشکی آشنا باشید.

تفاوت علوم انسانی و علوم طبیعی

توجه به این نکته ضروری است که در علوم انسانی (humanities) برخلاف علوم طبیعی(natural science) تعریف واحد و سرراست و نظریه‌ای فراگیر و جامع وجود ندارد، بلکه در هر موضوعی مکاتب، نظریه‌ها، دیدگاه‌ها، رویکردها و تعاریف مختلفی وجود دارد. این ویژگی در علوم انسانی *نه به معنای تشتت و سردرگمی است و نه به معنای نقصان دانش آن*، بلکه به سرشت این علوم بازمی‌گردد . مکاتب، نظریه‌ها، دیدگاه‌ها در علوم انسانی همچون زبانهای مختلف هستند که تلاش برای از بین بردن آنها و ساختن یک زبان واحد نه ممکن است نه معقول .در هر نظریه یا رویکردی در مقایسه با سایر نظریه‌ها و رویکردها، زمینه‌ها، پیش‌فرض‌ها، مسائل مرکزی، مفاهیم کلیدی، نحوهٔ پرداختن به مسائل و استدلالات و براهین متفاوت است.

شیوهٔ خواندن متون در علوم انسانی

منابعی که پیش رو دارید در تلاش خواهد بود که مطالب مورد نیاز را برای این مهم فراهم کند. از آنجا که متون برگزیده از نویسندگان مختلفی است تبعاً لحن‌های متفاوتی دارند و مسلماً یکدست نیستند، اما اگر پرسش‌های اصلی و کلیدی‌ای در خواندن متون مدنظرتان باشد مطالعه هدفمند و مؤثر رخ خواهد داد. با توجه به عدم وجود مطالب مشابه در برنامه‌های رسمی دانشجویان رشته‌های مختلف علوم پزشکی در این مورد ممکن است نیاز داشته باشید برای فهم برخی از مطالب به منابع کمکی مراجعه کنید. به هر رو هدف کلی نه به خاطر سپردن مطالب بلکه درک و فهم ژرف آنها برای به کار بستن آنها در تحلیل و نقد خلاقانه و بدیع است.

پرسش‌های کلیدی برای خواندن متون

در خواندن این متون انتظار می‌رود دانشجویان عزیز:

اول- مفاهیم اصلی MH بشناسند و رویکردهای مختلف در این حوزه را بازشناسند و تحلیل کنند.

دوم- آنچه از مفاهیم و رویکردهای مختلفMH فراگرفته‌اند را در تحلیل و ارزیابی انتقادی روشمند تکنولوژی هوش‌مصنوعی در حوزۀ سلامت به کاربندند.

سوم- تعاریف و رویکردهای مختلف به فلسفه تکنولوژی پزشکی را بدانند.

چهارم- انتظار می‌رود مفاهیم و رویکردهای نظری به تکنولوژی در حوزۀ سلامت بشناسند و آن را در نقد هوش‌مصنوعی در سلامت به‌کارگیرند.

پنجم- هوش‌مصنوعی در حوزۀ سلامت و چالش‌ها و پیامدهای فرهنگی، اجتماعی، حقوقی، اخلاقی و مسائلی ازاین‌دست را بازشناسند.

مطالعۀ تحلیلی و ژرف

از شما انتظار می‌رود نه‌تنها در خواندن متون تلاش کنید هر نظریه یا رویکرد را با توجه به محورهای بالا فرابگیرید، بلکه در برخورد با هر مسئله با مداقّه و آوردن براهین نشان دهید که کدام رویکرد یا نظریه برای مواجهه با آن مسئله مناسب‌تر است، چگونه می‌توان آن را در مواجهه با مسئله بکار بست و با بکار بستن نظریه مسئله را چگونه می‌توان صورت‌بندی کرد. در ضمن دانستن اینکه هر نظریه یا مفهوم را چه متفکری پیشنهاد کرده است بخش مهمی از مباحث علوم انسانی به همین دلیل در مطالعۀ متون این موضوع مدنظرتان باشد.

ساختار منابع آزمون

- منابع آزمون مرحلۀ غربالگری (ص ۹-۹۷) در سه بخش تنظیم‌شده‌اند. بخش اول اصول و مبانی علوم انسانی پزشکی است که شامل ۳ مقاله است. بخش دوم مفاهیم اساسی فلسفۀ تکنولوژی پزشکی است که سه متن به آن اختصاص‌یافته است. در بخش سوم که مرتبط با صورت‌بندی و چالش‌های هوش‌مصنوعی در حوزۀ پزشکی و مراقبت که از ژورنال‌های معتبر پزشکی آورده شده است.
- منابع آزمون مرحلۀ دوم انفرادی و گروهی (ص ۹۹ تا ۱۵۶) به تحلیل هوش‌مصنوعی پرداخته است.

***** از صفحه ۹ تا ۹۷ منابع آزمون غربالگری (مرحله اول انفرادی) است.**

***** از ص ۹۹ تا ۱۵٦ منابع صرفاً مربوط به آزمون انفرادی مرحلۀ دوم و آزمون‌های گروهی است و در آزمون غربالگری (انفرادی مرحلۀ اول) از این بخش‌ها سؤالی نخواهد آمد.**

***** از ص ۹ تا ۹۷ در آزمون‌های انفرادی مرحلۀ دوم و آزمون‌های گروهی سوال طرح خواهدشد.**

منابع برای مطالعهٔ بیشتر

ازآنجاکه دانشجویان رشته‌های علوم پزشکی عموماً دانش و زمینهٔ کافی دربارهٔ بحث‌های علوم انسانی را ندارند و نیاز است تا در مورد کلیات هم دانشی کسب کنند، منابع زیر برای مطالعهٔ بیشتر معرفی می‌گردد. بدیهی است در آزمون‌های المپیاد از این منابع سؤالی طرح **نخواهد** شد. مطالعهٔ این متون به شما کمک می‌کند، منابع اصلی را بهتر دریابید و چگونه در تحلیل و ارزیابی انتقادی هوش‌مصنوعی از چارچوب‌های علوم انسانی به شکل روشمند و آکادمیک بهره بگیرید.

- جعبه‌ابزار فیلسوف- برگزیده‌ای از مفاهیم و روش‌های فلسفی، نویسندگان جولیان بگینی و پیتر فوسل
- مثل فیلسوف نوشتن (آموزش نگارش مقاله فلسفی) نوشته الویشس مارتینیچ
- درآمدی بر فلسفهٔ تکنولوژی- وال دوسک
- تکنولوژی: فرانکنشتاین یا پرومته؟ غلامحسین مقدم حیدری و علیرضا منجمی

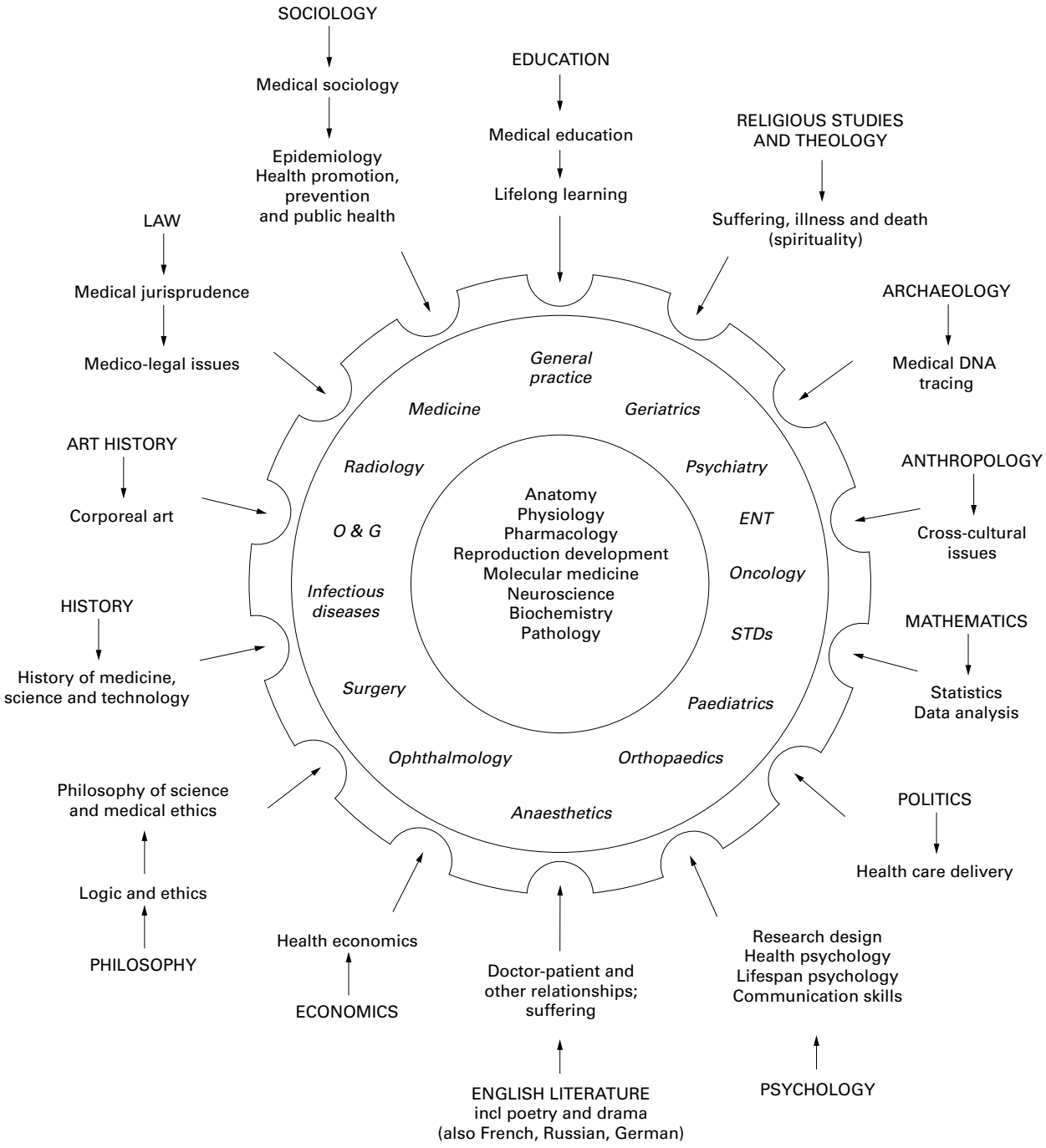
مؤخره

لازم به ذکر است کمیتهٔ علمی، این حیطه را چنان طراحی کرده است که نیل به این اهداف در فرآیندی چندماهه میسر است که هم تلاش شما را می‌طلبد و هم کمیته علمی با فراهم آوردن فرصت‌های آموزشی -در قالب وبینار و کارگاه -در غنی‌سازی و ارتقاء آن می‌کوشد. المپیاد فرصتی فراهم می‌کند تا در کنار یکدیگر یکی از حوزه‌های کمتر شناخته‌شده در حوزهٔ سلامت را بشناسیم و دربارهٔ آن با یکدیگر بیندیشیم. بی‌تردید مسابقه انگاری و فروکاستن آن به چند آزمون و توزیع چند مدال این کوشش علمی را مخدوش خواهد کرد و هدف نهایی از برگزاری المپیاد هم این نیست.

با آرزوی موفقیت برای شما دانشجویان عزیز و به امید آیندهٔ پررونق مباحث MH به همتتان.

کمیتهٔ علمی حیطهٔ مطالعات میان‌رشته‌ای علوم انسانی و سلامت
سیزدهمین المپیاد علمی دانشجویان علوم پزشکی کشور

Downloaded from http://mh.bmj.com/ on June 13, 2015 - Published by group.bmj.com



The Field of Medical Humanities

Part I
Medical Humanities:

Basic Concepts and Approaches

.متن کامل مقاله را از سایت مجله رایگان دریافت کنید

فلسفهٔ علم، پژوهشگاه علوم انسانی و مطالعات فرهنگی
دوفصل‌نامهٔ علمی (مقالهٔ علمی ـ پژوهشی)، سال دهم، شمارهٔ دوم، پاییز و زمستان ۱۳۹۹، ۲۲۵–۲۴۹

علوم انسانی پزشکی/ سلامت:
تحلیل انتقادی مبانی نظری و عملی پزشکی

علیرضا منجّمی*

حمیدرضا نمازی**

چکیده

«علوم‌انسانی پزشکی» Medical Humanities در وهلهٔ اول عبارتی نامأنوس به نظر می‌رسد. اینکه چگونه دو حوزهٔ مجزا و متمایز معرفتی همنشین شده‌اند، به وضعیت پروبلماتیک پزشکی اشاره دارد. در بخش ابتدایی مقاله به تحلیل علوم‌انسانی پزشکی بر اساس مناقشه‌های این حوزه خواهیم پرداخت و در بخش دوم حوزهٔ انتقادی مطالعات فراپزشکی به عنوان بدیل علوم انسانی پزشکی پیشنهاد خواهد شد.

جریان معاصر علوم انسانی پزشکی با نقد پزشکی مدرن از اواخر دههٔ شصت و اوایل دههٔ هفتاد میلادی آغاز شد که دغدغهٔ گسترش روزافزون علوم زیست‌پزشکی و انسان‌زدایی پزشکی را داشت. با مرور و تحلیل دقیق پژوهش‌ها و ادبیات علوم‌انسانی پزشکی پنج مناقشه اصلی شناسایی شدند: تعابیر و تعاریف گسترده و مختلف، رشته – حوزه، چندرشتگی – میان‌رشتگی، علوم انسانی پزشکی – علوم انسانی سلامت، علوم انسانی کلاسیک – علوم انسانی انتقادی و علوم انسانی پزشکی – فلسفه پزشکی. در تحلیل نهایی در لایه زیربنا دو عنصر را می‌توان از هم بازشناخت: یکی دوگانه‌هایی و دیگری رانه‌ها یا فرآیندها. دوگانه‌ها را می‌توان ذیل چند گروه کلی‌تر دسته‌بندی کرد:

---

* استادیار گروه فلسفهٔ علم و فناوری، پژوهشکدهٔ مطالعات فلسفی و تاریخ علم، پژوهشگاه علوم انسانی و مطالعات فرهنگی، تهران، ایران، monajemi.alireza@gmail.com

** استادیارگروه اخلاق پزشکی، دانشکدهٔ پزشکی، دانشگاه علوم پزشکی تهران، تهران، ایران (نویسندهٔ مسئول)، hrnamazi@tums.ac.ir

تاریخ دریافت:۱۳۹۹/۰۳/۱۱ تاریخ پذیرش:۱۳۹۹/۰۶/۱۰

Copyright © 2018, IHCS (Institute for Humanities and Cultural Studies). This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International, which permits others to download this work, share it with others and Adapt the material for any purpose

Medicine, Health Care and Philosophy (2007) 10:363–372
DOI 10.1007/s11019-007-9079-x
© Springer 2007

Scientific Contribution

Medical humanities: stranger at the gate, or long-lost friend?

H. M. Evans
Centre for Arts and Humanities in Health and Medicine, Durham University, St Hild's Lane, Durham DH1 1SZ, United Kingdom (Phone: +44-191-334-8245; E-mail: h.m.evans@durham.ac.uk)

Abstract. ''Medical humanities'' is a phrase whose currency is wider than its agreed meaning or denotation. What sort of study is it, and what is its relation to the study of philosophy of medicine? This paper briefly reviews the origins of the current flowering of interest and activity in studies that are collectively called ''medical humanities'', and presents an account of its nature and central enquiries in which philosophical questions are unashamedly central. In the process this paper argues that the field of enquiry is well-conceived as being philosophical in character, and as having philosophy – albeit pursued over a larger canvas – at the core of its contributing humanities disciplines. The paper characterises humanities disciplines as having an important focus on human experience and subjectivity, of which the experiences and subjectivities at stake in health, medicine and illness form an important sub-set, the preoccupation of the medical humanities as a whole. Claims of interdisciplinarity (as distinct from multidisciplinarity) are noted, but such claims need to be recognised for the high and stern ambition that they embody, and should not be made lightly.

Key words: humanities, interdisciplinarity, medical humanities, philosophy of medicine, subjectivity

In *The Curious Incident of the Dog in the Night-Time*, author Mark Haddon (2002) describes a fairly ordinary sequence of domestic unhappiness through the utterly extraordinary eyes of a logically clever, but emotionally severely disabled, teenager suffering from a form of Asperger's or other quasi-autistic disorder. The result of his condition is a quite unforgettable re-ordering of the world into bizarre yet internally consistent categories, including what is for the reader a heartbreaking systematic misperception of parental love as murderous threat; the book is a chronicle of how so disabled a child can somehow craft his own day-to-day survival. After reading this book I asked an experienced child psychiatrist whether he felt that the author had succeeded in capturing the ''interior'' of an autistic or Asperger child's experience. His answer was: ''not quite'', but that even with its inaccuracies he remained very glad that the book had been written, because in his view it made available the intensity of the problems of Asperger's and autism to a wide audience, and would generate sympathy and understanding of the condition. (I will from now on use the terms ''autism'' or ''autistic'' as an un-scientific shorthand to cover the range of Asperger's-like and other autistic conditions in general. The points I wish to make do not depend on the distinctions between these terms.)

The psychiatrist's answer – that the book had ''*not quite* succeeded'' – is an interesting one, for it implies the possibility of success. This in turn implies a number of moderately striking things, among them that, with sufficient professional experience, it is possible for the clinician to gain genuine insight into the interior of someone else's experience even in such notoriously inaccessible conditions as autism. That assumption is implicit in his being able to give a cautious approval of the book's partial accuracy – if I may use the term – i.e. partial accuracy with respect to a strange (and, in this particular case, damaged) form of self-experience and self-understanding. Of course this is an unusually difficult form of something that is somewhat difficult in even an *ordinary* case – namely, to get a sufficient degree of access to

someone else's experience, through what they write or say about themselves, for *us* to be able to talk about how successfully they have conveyed their experience, or how accurately they have represented it. The familiar obstacle is (depending on your position within the philosophy of mind) that since anyone's own experience is something that only he or she actually has, it can never be more than *inferred* by third-parties, that is, everyone other than that person.

However, the attempt to infer it – in the ordinary case – is obviously necessary a thousand times a day; and presumably it is no less necessary in trying to understand the perplexing case of autistic experiences. The psychiatrist's answer presumes this, too. He could hardly try to work clinically with autistic children and their parents if he had *no* ambition to understand something of the qualitative reality of autistic experiences, since without such understanding, the clinical role becomes reduced to something like advising the affected parents on the practicalities of crisis management.

The further implication of this verdict of partial accuracy is the possibility that the book could have succeeded in transmitting experience *among* third-parties: that is, an originating third-party such as a well-informed author could not only access such an obscure experience but also convey it accurately to other third-parties, namely ourselves as readers.

A somewhat distinct presumption in the psychiatrist's stance is the value of wider sympathy and understanding of the condition of autism. However intuitive such a presumption may appear, there is a question about where exactly that value lies. Would we be happy, for instance, if managerial decisions about funding and resources were openly based upon the extent to which the book-reading public sympathised with the plight of a particular group of patients and their carers? Surely not. Perhaps instead it lies in the likelihood that readers of the book will be more tolerant of the problems caused by autistic behaviour – and more supportive of the parents who routinely deal with those problems. But even this is problematic, in that whilst tolerance *per se* seems to be a good thing, we surely want it to be based upon a genuine and honest understanding, and not upon an inaccurate, picturesque, imagined or otherwise deficient representation. This seems to require, in the present case, that the book actually succeed in opening a genuine window into the autistic child's world. "Not quite" succeeding, in the psychiatrist's words, seems to be an imperfect basis for greater tolerance.[1]

The reason I have opened with this example, and spent some time on it, is that it raises a number of questions with which I think the fledgling field of the medical humanities is concerned. Let me briefly review a list of the more obvious of these questions. First, how far is clinical medicine based upon scientific observation and intervention? What resources other than scientific observation and intervention are available to the clinician?[2] Is clinical medicine directly, or only indirectly, concerned with the experiential aspects of health and illness? In either case, how do we train doctors and other clinicians to address these experiential aspects (and hence do doctors need experience of life, as much as they need scientific knowledge, in their clinical practice)? How should we seek to understand and explore those problems of life and experience, including physical and psychological illnesses, that arise from the particular configurations of our bodily make-up? The suffering of any illness, not just the suffering of ingrained emotional deficits associated with some psychological disorders, is an intensely subjective matter. What kinds of knowledge and enquiry therefore are best suited to taking subjectivity seriously, and investigating it? Perhaps more radically, can there really be genuine knowledge of another person's subjectivity? And if there can, how is it to be achieved? Furthermore, how can it be usefully *transferred* – for instance, making an exploration of the autistic child's subjectivity a matter for a gain in the understanding of others?

Other epistemic questions as well are implicit in the psychiatrist's answer. What is the role of *values*, be they moral, social, aesthetic and so on, in our identification of the normal and the pathological? What kind of science-of-the-human is constituted by medicine in either its early modern form or its current, highly technologised form? Is it central or peripheral in the context of other sciences? How are we to consider a form of scientific *object of study* – the patient – that is also a thinking, experiencing *subject*? How should we understand such a science in a context that is increasingly dominated in an epistemic sense by, on the one hand, biophysical categories including those of molecular genetics, and on the other hand, statistics and the relationship between probabilities at a population level and the individual patient – who supplies, perhaps, the *only* context in which these questions are finally important? And so on.

All of these questions present, constitute, or point towards, problems and enquiries that are recognisable in the philosophy of medicine, and I acknowledge the need to clarify the relations between philosophy of medicine and the field of medical humanities. But the fact that these questions are indeed recognisable in itself suggests that from the standpoint of philosophy of medicine, medical humanities begins to look a little more like a long-lost friend than simply a stranger at the gate.[3]

To continue the enquiry, I will try to present an individual perspective upon the field's origins and its contemporary nature. This perspective involves the frequent occurrence of irreducibly philosophical questions; in this paper I can only notice them and not address them substantively.

Origins

To begin with the field's origins, it is perhaps worth noting that the expression "medical humanities" is initially an American one, referring to primarily education concerns within the medical curriculum, and more particularly to bringing the study of humanities topics, principally literature and literary techniques, to the *teaching* of medicine; part of the aim was to develop clinicians' powers of listening and interpretation (Hunter et al., 1995). One difficulty the expression presents is that one always has to explain that "medicine" means other aspects of health care as well.

Another difficulty – and this implicitly brings us to the question of the field's current nature – is that some people regard medical humanities as of interest only within medical education, and indeed as essentially *being* a mode of medical education. So, to the extent that they are engaged at all in medical humanities, British medical schools have tended to maintain the original American approach of focusing on such things as literature, creative writing and film as vehicles for interpretation skills and self-expression alike. One development of this in the UK focuses on postgraduate and continuing education, using familiarity with the humanities and creative arts as a personal resource for hard-pressed clinicians facing the demands of professional life.[4] Another development emphasises the cathartic benefits to *patients* as well as carers, of writing creatively about their experiences (Bolton, 2001). These resources are no doubt all good things to have, but they do not in themselves plausibly constitute a field of study.

A further suggestion that has been made during the early evolution of medical humanities is that it is the attention we pay to (in the British sense) all the non-scientific (though not unscientific) aspects of medicine, or even simply all that concerns "the human" within medicine.[5] The trouble with this suggestion is that it is so dismayingly wide that it would be difficult to see how it could possibly be the name of a coherent activity or enquiry.

There is also a sense that medical humanities is a kind of *medical counter-culture*: a response to some forms of dissatisfaction felt by patients concerning how well their doctors listen to them, or dissatisfaction felt by doctors towards the somewhat dehumanising effects of large-scale, industrialised health care (Macnaughton, 2001). In this sense, such dissatisfactions (and they are not unreasonable) rather resemble some of the origins of medical *ethics* – that is, a kind of consumer voice of protest, *seeking* a critical counter-culture of this kind. This in turn invites a further resemblance to some of the early critical enthusiasm for medical ethics, before it met the twin dangers of being either turned into a respectable academic discipline or devoured by the law and legalism.

Perhaps this is a good point at which to review other parallels between medical ethics and medical humanities. "Medical ethics" is an ambiguous phrase with at least two denotations: on the one hand sets of practical and professional duties and their consequences (i.e. what actual, particular doctors ought to do in real situations, conveniently dubbed "normative medical ethics") and on the other hand a set of intellectual questions and enquiries which have been collected together as an academic field (i.e. how we might think about and understand what doctors in general ought to do in typical situations, sometimes called "critical medical ethics"). Now we might at first glance suppose that the phrase "medical humanities" is ambiguous in the same way.

I have discussed this problem at greater length elsewhere, and here I will merely summarise that discussion. The phrase is ambiguous between a form of enquiry and an approach to practice. The former denotes a specific branch, particularly aimed at medicine, of the broader area of enquiry known as "the humanities"; this critical and reflective undertaking corresponds to the critical form of medical ethics. (Such enquiry naturally includes questions in metaphysics and epistemology, insofar as these are natural components of any genuinely critically reflective examination – such as philosophical examination, whose importance we

shall recognise below – of medical practice and medical theory, which inevitably presuppose some metaphysical and epistemological positions on matters concerning, respectively, the nature of embodied human experience in health and illness and the sources of our knowledge of such experience and its bodily foundations.[6]) It contrasts with the advocacy of particular ways of actually doing medicine, that is, practising humanely and with due concern for the humanity of the patient; this exhortatory discourse corresponds to normative medical ethics. Unfortunately the problem for this latter interpretation of "medical humanities" is that it appears suspiciously like a truism of a rather pious kind.

It would certainly be a truism if humane practice were intrinsic to the concept of medical practice. However, this can be contested – as can the somewhat parallel presumption that ethical practice (of which humane practice might be thought to be a manifestation) is internal to medicine. In taking the relief of suffering as being an *internal* goal of medicine, Cassell (1991), for instance, seems committed to the idea that medicine in practice must be both ethical and humane *by definition*, a view whose consequence would be that if we fail to practise medicine humanely or ethically we fail to do medicine at all rather than just doing medicine badly, which seems on the face of it the more natural way of putting the matter.

If, prompted by caution, we disregard the normative interpretation of "medical humanities" as referring to particular (humane) ways of doing medicine then we are left with the still-valuable denotation of a critically reflective field of intellectual enquiry, and in this too, there is a useful parallel with medical ethics. I find persuasive the suggestion that medical ethics' concerns can themselves be taken up amongst the "human" (*not* humane, be it noted) concerns of medicine. In this sense, medical humanities adopts part of the agenda of medical ethics but pursues it in a broader and perhaps more diffuse form.

Of course "ethics" is the specific name of only *one* area of values, and there are other areas that are at stake in modern medicine and healthcare – social values, political values, spiritual values, aesthetic values, epistemic values, perhaps sexual or gender values, even gustatory values. Despite their obvious relevance to clinical medicine (think of public health, palliative care, aesthetic and reconstructive surgery, the fashionable preference for population-level evidence and so on), some of these have received relatively little attention, and I have a sense that this reflects a wider neglect of the philosophy of medicine – at least in the UK where, it seems to me, most philosophy of medicine is done in conjunction with medical ethics, perhaps actually as *part of* medical ethics. That is a legitimate place to do philosophy of medicine, of course, since critically undertaken value enquiry with regard to medicine is as fully a part of philosophy of medicine as is the pursuit of any of the other cardinal components of philosophical enquiry – epistemology, logic, metaphysics and so forth – directed at our understanding of medicine, whether undertaken in an analytic or an interpretative spirit.[7] Moreover from the philosopher's viewpoint it is an enduring boon that medical ethics has provided this opportunity, since through its relatively high profile medical ethics makes some philosophical questions apparent, and even accessible, to a wider public. Medical ethics is, as one might put it, the most brightly illuminated shop window display of *any* form of philosophical enquiry.

Before we leave the question of the origins of medical humanities, it is worth including a cautionary note (one which may be somewhat familiar in medical ethics as well), namely that the very imprecision – so far – of what medical humanities comprises, can appear to offer a home for what one might call disciplinary refugees: that is, enquirers who for one reason or another are not comfortable within the traditional confines of their own discipline or practice, and have moved into the area of reflective enquiry into medicine, hoping to claim the academic equivalent of political asylum. The benefits of intellectual creativity that such a diversity of individuals in theory offers may be offset by the adverse impact of too many varying influences upon a field of enquiry that is not yet itself sufficiently mature to be entirely confident of its own general nature, still less its detailed identity and purposes.

Unfortunately amidst a clamour of voices, one has rarely the luxury of waiting for silence before adding one's own voice. All I can therefore do in the remainder of this paper is offer a personal contribution to the discussion of the nature of medical humanities as a field of enquiry.

Nature

In the personal conception of the nature of the field of medical humanities which now follows, I will try to begin *descriptively*, reporting on what I see when

I look at the field, whilst acknowledging that the report inevitably involves a somewhat editorial selection on my part and, as such, is liable to develop *prescriptively*, advocating a particular conception.

The simplest pattern that I can impose upon a varied field of activities claiming to constitute, or at least to affiliate to, medical humanities is to divide those activities essentially into three kinds. The first two kinds concern *substantive activities* within organised health-care, as well as academic or theoretical *reflections upon* those activities.

- First, there are those activities collectively known as "Arts in Health" including the therapeutic uses of creative arts activities such as writing and painting; and including also the use of creative arts and co-operative productions of public art as a way of helping to create and sustain healthy communities. An example of the former would be the encouragement of creative writing on the part of sufferers of chronic illness – or their carers – in an attempt to confront and give meaning to symptoms (Bolton, 2001). An example of the latter would be the use of stylised visual rituals, such as the lantern project in Wrekenton, near Durham in the North East of England, in which illuminated symbols of the heart at the core of a healthy community are produced collectively in community-based workshops and then paraded together in an annual and spectacular festival of lanterns (Robson and White, 2003). As mentioned, for me this area of medical humanities includes commentary, analysis and critical reflection upon arts-in-health activities.
- Second, there are those activities geared towards and embedded within Medical Education, including actual schemes of study for medical undergraduates and postgraduates, periodic study resources for Continuing Medical Education, and the general notion of offering personal resources, through art, literature and creative self-expression, for what I earlier referred to as "hard-pressed clinicians facing the demands of professional life". Examples of modules devoted to the study of literature, film, fine arts, history and philosophy can be found in many medical schools, normally as options,[8] and as part of continuing medical education through, for instance, the Medical Royal Colleges in the UK.[9] Again this area should be taken to include academic commentary and analysis concerning such activities.
- The third area is more obviously an academic or theoretical undertaking *through and through* – namely, the task of attempting better to *understand human nature* through the lens of a critical examination of technological medicine and its limitations. Examples of enquiries here could include the implications of molecular genetics for our concepts of free will; scrutiny of the role of technology in medicine in an age in which imaging the body's interior is taken to have category-forming authority and explanatory power (Hofmann, 2001); or the two-way relationship between new surgical techniques and contemporary standards for so-called "ideal" bodies and faces.[10] This is not only the most clearly theoretical of the three broad areas of work; it is also the most irreducibly philosophical of the three. Whilst I do not want to suggest for a moment that only philosophers can undertake it, I do want to suggest that in undertaking it one is doing something that, whatever else it is, is usually also philosophical in spirit.

So, if we try to identify the nature of the medical humanities in terms of its characteristic preoccupations, then these three broad areas seem to me to describe it. But an equally important question concerns *who is* actually so preoccupied: Which *are* the contributing disciplines to the field? Well, almost by definition they are neither physical sciences nor, for the most part, social sciences. No doubt the division of human enquiry into discrete disciplines is a historical and conventional one that is in some respects unhelpful, but we are stuck with it and we might as well start from where we find ourselves. So, we are left with the humanities disciplines, whose conventional members include literature studies, history, philosophy, fine art, drama, critical theory, historiography, theology and religious studies, linguistics, music, law and so forth. The least generalising of the social sciences (the qualification is important as we shall shortly see) such as ethnography or that borderline humanities/sciences discipline, psychology, might also be included in an eclectic conception.

A putative list is all very well – although of course people will disagree over the inclusion of some of these, and over the exclusion of a larger number of disciplines not mentioned here (how about cultural anthropology or feminism studies?) – but we need to go on from this to ask, Do they have anything in common that makes them either characteristically *humanities* disciplines or specially able to contribute to medical humanities study? I will try to respond to this by suggesting that there are indeed two related characteristics of humanities disciplines that do make them especially useful for addressing the human side of medicine.

These are, first, a concern with experience – with recording and understanding and interpreting individual human experience (Evans, 2002b) and its qualitative dimensions, or, if you like, a concern with the world as it is humanly encountered, rather than as it might be detached and merely dispassionately observed, which is more plausibly the goal of the natural sciences.

The second characteristic of the humanities for me follows from this (at least in broadly Western culture where, currently, conventional humanities subjects as characterised above, and the medical humanities as a manifestation of them, are primarily to be found). This second characteristic is a concern to take subjectivity seriously – the individual point of view and its qualitative content, its unique antecedents and its idiosyncratic repertoire of meanings and connotations – as well as taking seriously its necessary reflection of, and embeddedness in, the many *interpersonal* contexts of society, including those of clinical medicine.

This second characteristic invites us to suppose that the specific observations of a given individual in context are as interesting – in the sense of providing grounding, meaning, implication and a guide to our future attitudes and actions in relevantly similar circumstances – as are the homogenised observations collected together under the natural sciences. It allows that for many purposes characteristic of clinical practice (such as the decision of whether or not to prescribe a marginally effective drug with unwanted side-effects), a single telling example of a vivid experience that is to some extent recognisable to us is, in principle, as powerful as population-derived evidence telling us which probabilities are compelling as guides to action (Sweeney, 1996).

The point is that both the objectivising gaze of science and what we may call the subjective-tolerant gaze of the humanities do indeed contribute to our reasoning as guides to future action. I should like to attempt a generalisation here – a generalisation that, if plausible, helps to rationalise the place of the humanities in our understanding of medicine, health and illness: perhaps the sciences provide constraints upon what is a *rational conception* of future action – they provide the basis for our beliefs. At the same time, perhaps the humanities provide models of *motives* for future action – they provide possible bases for our attitudes (what Stuart Hampshire (1989) called our conceptions of the good lives that are possible for us).

Having suggested the broad content of the field of medical humanities, and characterised the humanities disciplines that engage in it, I would like to add something about the *modus operandi* that is at any rate claimed for Medical Humanities. This is its alleged interdisciplinarity. Most promotional references to medical humanities advertise this as a characteristic feature. However, I suggest that interdisciplinarity is a very ambitious goal, and that it is claimed on many more occasions than it is actually realised. This is arguably a further feature in respect of which medical ethics and medical humanities are somewhat alike.

First, however, what is at stake in attaining a proper conception of ''interdisciplinarity''? Principally at stake is the way in which the various contributing disciplines are thought to relate to one another as they jointly engage medicine and health care. How do these actually *constitute* medical humanities as a field of enquiry?

The essential question here is whether the contributing disciplines remain as independent of one another as, inevitably, they must begin. For example, the question of the status of neurasthenia (in some respects, the late-19th-century counterpart of myalgic encephalopathy) as a genuine disease invites commentary from history of medicine (in terms of the emergence and refinement of an identifiable condition attracting medical attention), literature studies (in terms of the coalescing of references to the condition around certain prominent artistic or creative individuals at a particular historical period, and the value-assumptions that began to be tied to the condition) and philosophy (in terms of genesis and maturation of the concept ''disease''). The question is whether these several enquiries are, or could be, or should be, undertaken in radical independence of each other; or, if not, the contrapuntal question is that of precisely how they should inform one another. Are they a mere sequence of set-piece investigations to be sampled piecemeal according to the interests of the external inquisitor, or are they the fused components of a more richly-layered and above all *emergent* enquiry, whose substance, concerns and specific questions would not be apparent to the contributing disciplines on their own?

This is of course a puzzle about what kinds of knowledge are possible when distinct disciplines collide, about whether their respective methods are mutually intelligible, about ''how other disciplines see and name the objects in their world, and to what extent we can view that world with them: in effect, learning to see simultaneously through our own eyes and through theirs'' (Evans, 2007).

No doubt true interdisciplinarity is sometimes achieved, but so far the more convincing examples appear to emanate from elsewhere than the medical humanities. A good example is arguably constituted by chemical process technology, in which those who, for commercial reasons, were interested in improving the mechanics of fluid flow and heat transfer in the production of polymer plastics, had initially no established field to draw upon at all (Evans and Macnaughton, 2004). Proceeding empirically, they engaged mechanical engineers to help them with pencil and paper calculations; the engineers in turn recruited methods from physics involving so *many* simultaneous calculations that non-linear mathematical modelling from computing science became integral to the emerging field.

A key feature of this process is that at each stage new *questions* emerged that could not have been asked, let alone answered, by the contributing disciplines in isolation. I think it is plausible to suggest that *emergent questions*, whose range of aspects cannot be found in any single contributing discipline, are one indication that genuine interdisciplinarity has been achieved. The full complexity of fluid mechanics was neither soluble by *nor apparent to* the paper-and-pencil generation of chemical and mechanical engineers who began the field; the relation between on the one hand real fluids traversing real locations and on the hand mathematical representations of activity at notional and infinitesimally graded locations would at earlier stages have seemed arcane to both fluid mechanics and computer scientists.

It is I think more difficult to point to either emergent questions in particular or genuine interdisciplinarity as a more general attainment within medical humanities at the moment. The mutual implications, for our understanding of perception, between neurology and phenomenological philosophy become apparent and real only when these two forms of enquiries collide. More generally, patients' subjective experiences are foundational in their seeking medical care in the first place, yet the *forms* of experiences of the self occupy a surprising range; some forms are perhaps even made possible when disciplinary perspectives co-engage. Consider, for instance, Oliver Sacks' incorporation of the notions of music and musicality into his understanding of proprioception as a neurologist, an incorporation that informs his experience of his own bodily recovery and our appreciation of music's diagnostic and therapeutic possibilities (Sacks, 1986, pp. 108–110). As for interdisciplinarity as such, one place where one might look for interdisciplinarity is where the methods of literary and philosophical analyses are combined – as has been fruitful in medical ethics and indeed ethics more generally. Examples might include the attempt to understand the processes of creative imagination in the evolution of scientific medicine, or the attempt to chart the complexities of paying attention to the character of the moral agent in expositions of virtue ethics. I am not here going to comment on the success or otherwise of any particular claim to interdisciplinarity. I merely want to insist on how difficult it is, at the same time as noticing how routinely and, I fear, *casually* it is claimed on behalf of Medical Humanities.

Notwithstanding this sceptical note, the foregoing (taken as a whole) suggests to me that we can say something about the characteristic projects of work likely to fall under the Medical Humanities. My suggestion is that at least such work as attempted any of the following four tasks – and it is straightaway apparent that they all have a philosophical flavour – could be thought of as constituting Medical Humanities work.[11] (That is to say, the attempt on these tasks provides a sufficient, although presumably not a necessary, condition for constituting Medical Humanities work.) The tasks are these:

1. To illuminate the practice of medicine (and, perhaps, medical theory) using ideas and insights distinctively associated with humanities or social science disciplines; especially doing so in a way that is not usually accessible through scientific descriptions and explanations.

   Examples: any kind of value enquiry concerning medicine. This obviously includes medical ethics.

2. To illuminate what one might call "the human side of medicine" in a form that takes seriously the ways in which medicine, illness, suffering, disability, and (for that matter) health are *experienced*

   Example: pathographies – the recording and interpretation of illness experiences; bringing creative and expressive arts to bear upon the experience of illness, in the therapeutic (and sometimes diagnostic) context

3. To attempt the understanding of one or more 'subjectivities' within the experience of medicine, or of health, illness, suffering or disability; and (from this) work that makes such understanding

transferable to our understanding of *other* subjectivities: such that we gain something which we can meaningfully relate to other insights gained on other occasions of comparable enquiry, allowing us to be systematic, albeit in a rudimentary way

> Examples: the broad swathe of those enquiries in history of medicine, philosophy of medicine or medicine and literature where individual experiences are made available to others through description, analysis, representation, in the hope of learning something about ourselves – and about "the human condition"

4. To use some aspect of medicine (that is, health care, etc.) specifically to achieve some gain in our understanding of the human condition, or of embodied human nature

> Example: philosophy of medicine generally, especially philosophical enquiries into embodiment and experience; or similar enquiries within medical anthropology and ethnography

What would be the point of the foregoing work? Why would we seek to undertake it? I put these questions somewhat rhetorically – since all of these kinds of work, especially the last area concerning gains in our understanding of embodied human nature, should commend themselves to all serious scholars and above all to philosophers. But rhetorical or not, we can I think see that work of this kind does help us to do a number of worthwhile things.

To begin with some fairly conventional objectives, the first three of these areas clearly help us – as commentators or as clinicians or, for that matter, as patients who necessarily contribute to the clinical consultation – to take human *values* seriously, including ethical values. They help clinicians and students alike to develop their own personal values. The second and third areas may help in developing clinicians' interpretative sensitivity and their skills of listening and communication. Through the engagement with creative and expressive arts, they may also indirectly provide clinicians with personal resources for facing the demands of clinical life.

The fourth of these areas – fittingly enough for work that is essentially philosophical – serves I think more radical goals:

- asking how technological medicine's picture of human nature/the human condition contributes to our self-understanding, and whether other pictures are available (for instance, from the humanities);
- from this, asking whether technological medicine spurs humanities disciplines to extend (or revisit) their own research agendas;
- exploring disciplinarity, interdisciplinarity, and the varying nature of knowledge and evidence in medicine, sciences and humanities
- stimulating and encouraging a sense of wonder at embodied human nature.

I believe all of these goals are worth pursuing. To varying extents, each of them is reflected in current work in philosophy of medicine. I would describe this area of Medical Humanities as, in effect, pursuing philosophical questions in medicine over a larger, a more colourful and no doubt a more disordered landscape. If I may so put it, the "long-lost friend" has indeed been a stranger, but at others' gates; it is returning now with tales of these colourful and disordered landscapes.

Finally, if the field is to develop credibly then, I would argue, its constitutive research enquiries must strive to be mutually coherent. Literary insights, historical investigations, philosophical reflections and linguistic analyses directed towards, say, culturally distinct experiences of nausea and their appropriate medical and psychological management (or towards the meaning of the epidemiology of psychological disorders, or towards the notion of "functional illnesses", or towards the question of whether myalgic encephalopathy is genuinely comparable to late-19th-century neurasthenia, or towards radical deconstruction of the clinical consultation, and so on) should be seen to bear upon common objects in compatible terms. Unfortunately I do not think we can always claim that this happens as yet. There needs to evolve at some point a rudimentary structure, within the field of Medical Humanities, that minimally orientates the modes of attention of different disciplinary enquiries and focuses them together upon an object or concept that is recognisable to all the enquirers and has a shared meaning as well as, putatively, *a shared denotation*. Research in medical humanities needs to produce some sense of accumulated gains in understanding, and not just an unstructured "heap" of observations and remarks that are individually valuable but nonetheless essentially fragmentary.

I do not suggest that this is easy, but few worthwhile things are easy. Elsewhere I have suggested that in the biomedical age we might recast Blake's powerful rendering of the human

constitution, famously the constitution of ''impassion'd clay'', in terms of our being ''meat with a point of view'' – the combined biophysical and existential realities of our embodied state, in which our subjectivity is fused with our objective, external being. Understanding this fusing is among the most philosophical of the tasks to which, in my view, the Medical Humanities are properly addressed.

This suggests that those who, as I do, prefer the ''long-lost friend'' conception of Medical Humanities to the ''stranger at the gates'', will recognise the centrality of philosophy among its contributory disciplines. Indeed I would go so far as to say that for those of its practitioners who are philosophers, the Medical Humanities amount to ''Philosophy looking at the Humanities looking at Medicine''. Further, the philosopher sympathetic to this view will sense that philosophy of medicine is the queen of those humanities disciplines co-engaging our embodied human nature. This is my sense, too. However, philosophy is not the only such discipline, and its task in the medical humanities is perhaps to encourage, to inspire, to learn from, to respect and, when necessary, politely to marshal the others. Whether this is finally a responsible and sustainable view, rather than unwarranted disciplinary arrogance, is something we shall find out only when the field of Medical Humanities progresses towards maturity.

Notes

1. *Perhaps* imperfectly grounded tolerance is better than nothing, if that is all we can get, but its wider consequences might involve more harm than good, if these include a more general decline in critical scrutiny of the bases of tolerance; we may end up tolerating things that we should *not* tolerate.
2. I am using the word ''scientific'' in its narrower UK sense. I mean by it the natural sciences, rather than the more general sense of organised knowledge implied by *Wissenschaft*, which extends to the humanities.
3. There are of course other viewpoints. Not all those engaged in clinical healthcare are so sympathetic to the programmes and projects of philosophy of medicine as to admit the value of medical humanities study through this particular door. I have elsewhere commended medical humanities to non-philosophical, expressly clinical, audiences; see for instance my 'Roles for Literature in Medical Education' (Evans, 2003); 'Reflections on the Humanities in Medical Education' (Evans, 2002b); or 'Medicine, Philosophy and the Medical Humanities' (Evans, 2002a).
4. The UK's first Master's in Medical Humanities, introduced in 1997 at University of Wales Swansea, appeals primarily to mid-career medical professionals. See Evans, M., in Kirklin and Richardson (2001).
5. Reported by Greaves (2001).
6. I am grateful to an anonymous referee for emphasising this.
7. The relation of philosophy of medicine to philosophy of science is an interesting one. Some enquiries within epistemology of medicine could readily be seen as an application of philosophy of science as could some enquiries within the logic of clinical reasoning and diagnosis. However, studies of the metaphysics of embodied experience will be more resistant to being captured in this way; indeed on Toulmin's (1993) view the centre of gravity of traditional views of philosophy of science is liable to be itself shifted by taking seriously the epistemology of medicine's objects.
8. See for instance Hampshire and Avery (2001).
9. The Royal College of General Practitioners' regional Faculties support specific study events involving medical humanities, and the Royal College of Physicians of London has published two volumes of papers on medical humanities including Kirklin and Richardson (2001).
10. Holm (2000). In 2005 the UK Arts and Humanities Research Council also sponsored a workshop at Univ. Cambridge on the human face, as one of a series of workshops exploring medical humanities enquiries.
11. Drawn from Evans (2007).

References

Bolton, G.: 2001, *Reflective Practice: Writing and Professional Development*. London: Paul Chapman.

Cassell, E.: 1991, *The Nature of Suffering and the Goals of Medicine*. Oxford: Oxford University Press.

Evans, M.: 2002a, 'Medicine, Philosophy and the Medical Humanities', *British Journal of General Practice* 52(479), 447–449.

Evans, M.: 2002b, 'Reflections on the Humanities in Medical Education', *Medical Education* 36(6), 508–513.

Evans, M.: 2003, 'Roles for Literature in Medical Education', *Advances in Psychiatric Treatment* 9, 380–386.

Evans, H.M.: 2007, 'Medical Humanities: An Overview', in: R. Ashcroft, H. Draper, A. Dawson and J. MacMillan (eds.), *Principles of Health Care Ethics*. (2nd.,). Chichester: John Wiley and Sons, 199–206.

Evans, H.M. and R.J. Macnaughton: 2004, 'Should Medical Humanities be a Multidisciplinary or an Interdisciplinary Study?', *Journal of Medical Ethics: Medical Humanities* 30(1), 1–4.

Greaves, D.: 2001, 'The Nature and Role of the Medical Humanities', in: M. Evans and I. Finlay (eds.), *Medical Humanities*. London: BMJ Books, pp. 1–22.

Haddon, M.: 2002, *The Curious Incident of the Dog in the Night-Time*. London: Jonathan Cape.

Hampshire, S.: 1989, *Innocence and Experience*. Cambridge: Harvard University Press.

Hampshire, A.J. and A.J. Avery: 2001, 'What can Students Learn from Studying Medicine in Literature?', *Medical Education* 35, 687–690.

Hofmann, B.: 2001, 'The Technological Invention of Disease', *Journal of Medical Ethics: Medical Humanities* 27(1), 10–19.

Holm, S.: 2000, 'Changes to Bodily Appearance: The Aesthetics of Deliberate Intervention', *Journal of Medical Ethics: Medical Humanities* 26(1), 43–48.

Hunter, K.M., R. Charon and J.L. Coulehan: 1995, 'Study of Literature in Medical Education', *Academic Medicine* 70(9), 787–794.

Kirklin, D. and R. Richardson (eds.): 2001, *Medical Humanities: A Practical Introduction*. London: Royal College of Physicians.

Macnaughton, J.: 2001, 'Why Medical Humanities Now?', in: M. Evans and I. Finlay (eds.), *Medical Humanities*. London: BMJ Books, pp. 187–203.

Robson, M. and M. White: 2003, 'Ice to Fire', Mailout, June 2003.

Sacks, O.: 1986, *A Leg to Stand on*. London: Pan Books.

Sweeney, K.: 1996, 'How can Evidence-Based Medicine Help Patients in General Practice?', *Family Practice* 13, 489–490.

Toulmin, S., 1993, 'Knowledge and Art in the Practice of Medicine: Clinical Judgement and Historical Reconstruction', in: C. Delkeskamp-Hayes and M.A. Gardell Cutter (eds.), *Science, Technology and the Art of Medicine*. Dordrecht: Kluwer Academic Publishers, pp. 231–249.

Medicine, Health Care and Philosophy (2007) 10:373–383
DOI 10.1007/s11019-007-9080-4
© Springer 2007

Scientific Contribution

Medical humanities and philosophy: Is the universe expanding or contracting?

William E. Stempsey
Department of Philosophy, College of the Holy Cross, One College Street, Worcester, MA, 01610, USA (Phone: +1-508-7932469; Fax: +1-508-7933841; E-mail: wstempsey@holycross.edu)

Abstract. The question of whether the universe is expanding or contracting serves as a model for current questions facing the medical humanities. The medical humanities might aptly be described as a metamedical multiverse encompassing many separate universes of discourse, the most prominent of which is probably bioethics. Bioethics, however, is increasingly developing into a new interdisciplinary discipline, and threatens to engulf the other medical humanities, robbing them of their own distinctive contributions to metamedicine. The philosophy of medicine considered as a distinct field of study has suffered as a result. Indeed, consensus on whether the philosophy of medicine even constitutes a legitimate field of study is lacking. This paper presents an argument for the importance of a broad conception of the philosophy of medicine and the central role it should play in organizing and interpreting the various fields of study that make up the metamedical multiverse.

Key words: academic disciplines, bioethics, medical humanities, models, philosophy of medicine

Introduction

Cosmologists debate the question of whether the universe is expanding or contracting. They have puzzled about an unresolved consequence of the big bang theory known as the flatness problem. At issue is how much matter there is in the universe. If the amount of matter is small enough, the universe will go on expanding forever. On the other hand, if there is a critical amount of matter, gravity will eventually stop the expansion and cause the universe to condense toward a "big crunch," possibly followed by a re-expansion. In the 1980s, Alan Guth developed his "inflation theory," which sees the origin of the universe in a tremendously rapid period of expansion in a tremendously short period of time, and there are now several versions of inflation theory. The one developed by Andrei Linde, known as the "bubble theory," proposes the possibility that other universes, presently unknown, might also have inflated, thus making our universe only one "bubble" in a much vaster "multiverse." While these "parallel universes" exist simultaneously, the finite nature of the speed of light makes it impossible for us to see into any of these other universes. Even in the midst of this explosion of theories, however, the question of whether our universe will continue to expand forever or collapse in a "big crunch" remains unanswered because we have no way to predict how much energy the universe contains (Siegfried, 2002, pp. 127–182).

I want to suggest that even though there are obvious limitations to the analogy, this image of a multiverse is an illuminating one for the present state of the medical humanities. "Medical humanities" is a term that is usually taken as a collective for various disciplines that study the human aspects of medicine, as opposed to the technical aspects. It includes such things as philosophy, theology, history, literature, and art, insofar as they are concerned with understanding medicine and medical practice. "Medical humanities" is also sometimes understood in a broader sense to include law, sociology, anthropology, and psychology. Work in the medical humanities seems to be expanding at present, but it is not at all certain whether this expansion will go on indefinitely or whether the enterprise will shrink or even collapse in upon itself. We just do not know how much energy there is in this academic world, and the data from which we might draw such conclusions at times seems as complex as the data from which cosmologists draw their speculation about the universe.

The medical humanities constitute a kind of academic multiverse, although it is a multiverse composed of the academic universes that are the traditional academic disciplines, and hence they interact more than the universes of Linde's bubble theory. What makes these universes cohere as a multiverse is that they share an appreciation of medicine as a human endeavor that reaches beyond its technical and scientific aspects. Their subject might aptly be called "metamedicine," which was the wonderfully descriptive and alliterative original title for the journal *Theoretical Medicine*, lately expanded to *Theoretical Medicine and Bioethics*. This titular evolution is, perhaps, a good indication that the "metamedical multiverse" is indeed expanding.

If we take the medical humanities to be a metamedical multiverse composed of the universes of philosophy, history, literature, etc. insofar as they concern themselves with medicine, there arises the question of how these various universes influence each other. I want to explore some models that describe these influences, and argue that the philosophy of medicine has a central role. Philosophy has always been the discipline that seeks the assumptions behind all human endeavors and the very essence of those endeavors; philosophy attempts to give an integrated account of these endeavors. Thus, philosophy of medicine seems the most likely candidate to serve as an integrating force in metamedicine. But we must also take note of a great gravitational force – some might say a black hole – that sometimes seems to be sucking many other metamedical studies, and even entire universes, into itself: bioethics. I will be particularly interested in the relationship of bioethics and the philosophy of medicine and the question of whether bioethics will ultimately doom philosophy of medicine to be lost in space.

Medical humanities

The most common understanding of medical humanities takes the field as an attempt to "humanize" scientific medical practice. David Greaves (2001, pp. 15–19), however, finds fault with most approaches to medical humanities because they maintain the traditional separation between medicine as an art and medicine as a science and side with the arts aspect to humanize the science aspect. Greaves (p. 22) distinguishes between medical arts, which attempt to humanize the physician, and medical humanities, which attempt to humanize medicine. He calls for a new conception of medical humanities that is humanistic in that it brings a "philosophical outlook" to both the science and the art of medicine. Greaves understands "philosophical" not in the restricted sense of philosophy as a field of study, but rather as an attitude of critical reflection. Medical humanities, then, promotes a humanistic perspective that attempts to unite the art and science of medicine.

This is a laudable goal, but what remains at issue is whether it is possible to conceive of medical humanities as a field unified enough to accomplish such a goal. Furthermore, we might well ask whether it is even desirable to conceive of medical humanities as an interdisciplinary field itself, and thus more than a metamedical multiverse of distinct academic universes reflecting on medicine. I have doubts about such conceptions, which will become more evident with some discussion of the notion of interdisciplinary fields and, in particular, bioethics.

Interdisciplinary and multidisciplinary fields of study

It is my contention that medical humanities do not constitute a field of study. Rather, "medical humanities" is a name given to the multiverse consisting of many academic universes that reflect on medicine, in both its theoretical and practical aspects. The medical humanities bring well-established disciplines such as philosophy, literature and history to a critical reflection on medicine.

This is not to say that the various fields that constitute the medical humanities are pure academic disciplines. For instance, the history of medicine is quite well established as a field of study, but it includes a disparate group of members, including both historians and physicians. The question of whether philosophy of medicine constitutes a distinct field has raised considerable controversy not only because it includes practitioners from both medicine and philosophy, but also because there is disagreement about exactly what subject matter constitutes the field.

Although medical humanities all attempt to lend a humanistic perspective to medicine, they do so in diverse ways. One doing a philosophical study of the logic of medical diagnosis, for example, approaches the task in a way that is very different from one studying a short story about a doctor puzzling about making a diagnosis that has important implications for a patient. Both shed light on the process of diagnosis, but the light comes from

quite different directions and is refracted in quite different directions.

That the medical humanities comprise many distinct academic disciplines and fields should not be seen as a liability, for this is precisely what makes the medical humanities such a rich human endeavor. It does, however, contribute to analytical complexity and controversy about how the parts relate to the whole.

When members of various disciplines meet to address topics of mutual interest, one might well ask how they see what they are doing. In *The Birth of Bioethics*, Albert Jonsen (1998, pp. 24–26) discusses the origin of the now superseded Society for Health and Human Values. The society was focused not only on ethical issues in medicine, but on the medical humanities, which included art, philosophy, history and literature. At the time it held its first annual meeting in 1970, it served as a meeting place for some "otherwise lonely figures," those few people who came from the disciplines of theology, philosophy, literature and art, and were now teaching in medical schools.

That society always struck me as multidisciplinary. That is, people from the various academic disciplines and the various health care professions came together to talk about their common interest – how to keep a human focus on an increasingly technological practice of medicine. Some people may have called themselves bioethicists because bioethics is what they did for most of the day, but they still identified in a more fundamental sense with their training as theologians, philosophers, physicians, nurses, etc. That sense of multidisciplinary cooperation is increasingly being supplanted by interdisciplinarity. Renée Fox and Judith Swazey (2005, p. 367) call bioethics "a multidisciplinary field with interdisciplinary aspirations." The distinction I am making here, which may not be exactly the same as that of Fox and Swazey, is this: a multidisciplinary endeavor is one in which people from several disciplines come together to talk about a topic of common interest. An interdisciplinary endeavor is one in which the endeavor itself is seen as growing from one comprising several distinct disciplines into a new "interdisciplinary discipline." In other words, multidisciplinarity is the meeting of people from different disciplines, who all retain their own sense of working in their own disciplines, while interdisciplinarity requires that each person be versed in several disciplines.

Evans and Macnaughton (2004, pp. 1–2) define a discipline as "a self conscious field of sustained, systematic inquiry with its own distinguishable subject matter, questions, and methods." Interdisciplinarity, then, is the engagement of disciplines with subject matter that "somehow both straddles the disciplines and falls between them." They suggest that the most important characteristic of interdisciplinarity is *emergence*. That is, particular problems and their solutions become evident, or emerge, only in the interaction of different disciplines, not within the disciplines by themselves. Furthermore, the participants that begin in different disciplines begin to share each other's metaphors.

My contention is that medical humanities is losing its multidisciplinary focus and moving more and more toward becoming interdisciplinary. This is coming about, I believe, because of the increasing acceptance of bioethics as a new discipline itself, an "interdisciplinary discipline." Bioethics, with its self-contained theoretical debates about such new ethical theories as "principlism," matters of informed consent arising from legal cases, and incorporation of principles such as double effect from moral theology, has provided a new *lingua franca* for medical humanities. Bioethics engulfs other disciplines, especially the philosophy of medicine, into itself. To see how this model has come to be so prominent, it will be helpful first to look at the development of bioethics as a new discipline.

Bioethics

Most observers trace the origins of bioethics back to about 1970. Of course, reflection on the ethics of medicine goes back at least to the time of Hippocrates, some quite specific ethical thought developed around medical issues in the Middle Ages, and medical ethics was developed systematically in the early nineteenth century, but present-day bioethics is seen to be different. Albert Jonsen (1998, pp. 3–33) finds the "birth of bioethics" rooted in the rapid changes in medicine following World War II. This prompted several conferences during the 1960s to reflect on the moral dimensions of these changes, followed by the establishment of two centers, the Hastings Center, outside of New York, and the Kennedy Institute of Ethics at Georgetown University in Washington. These centers provided a permanent home for discussions about the burgeoning questions of bioethics. A third organization, the previously mentioned Society for Health and Human Values, bolstered the development

of bioethics as a discipline by instituting a series of annual meetings of interested persons.

Warren Reich (1994, 1995) has argued that the word "bioethics" came into being independently at about the same time in two places, but with slightly different understandings. At the University of Wisconsin, Van Rensselaer Potter used the word to focus on a discipline that would study evolutionary and cultural adaptation in the context of the new biology in order to enrich human lives and prolong the survival of the human species. This conception of bioethics would embrace environmental concerns as well as medical ones. It was, in this sense, a holistic view. Potter regarded bioethics to be involved in "the search for wisdom," that is, for knowledge about what would enable good judgment about what was valuable for survival.

At Georgetown, on the other hand, André Hellegers was using the word to designate an academic discipline that would also focus on the interaction of science and ethics, but more narrowly on the realm of health care. The Georgetown model would seek to "resolve moral problems in three areas: (1) the rights and duties of patients and health professionals; (2) the rights and duties of research subjects and researchers; and (3) the formulation of public policy guidelines for clinical care and biomedical research" (Reich, 1995, p. 20). Reich (1995, p. 30) concludes that the word "bioethics" was what gave rise to the field of bioethics in part because "the word itself symbolized and stimulated an unprecedented interaction of biological, medical, technological, ethical, and social problems and methods of thinking."

Albert Jonsen (1998, pp. 327–342) argues that any discipline is characterized by the presence of a central theory, or sometimes alternative theories, principles, and a methodology to order, analyze, and evaluate the discipline's content. Bioethics has this to the extent that it has been formed into a body of knowledge that can be taught, and while it does have some elements of emerging theory, it is still not a discipline with any universally agreed upon methodology. As Jonsen (1998, pp. 342–344) says, bioethics is a "*mélange* of disciplines," including philosophy, theology, law, social sciences, and now more and more the arts and literature.

But Jonsen (1998, p. 346) has a further insight that is illuminating: he says that bioethics might well be considered a "demi-discipline." That is, only half of bioethics is like ordinary academic disciplines. The other half is a public discourse involving people of all sorts and professionals of all sorts arguing about bioethics, teaching it, and struggling to make practical decisions about how to deal with suffering. Bioethics, then, is a discipline unlike other purely academic disciplines, and more like a professional endeavor. From its earliest days, bioethics was shaped by the realization that its focus would be to help physicians to make hard decisions. It would have to move out of the ivory tower of academe and become as much a profession as an academic discipline. More than thirty years ago, Daniel Callahan (1973, p. 73) concluded his discussion of bioethics as a discipline: "The discipline of bioethics should be so designed, and its practitioners so trained, that it will directly – at whatever cost to disciplinary elegance – serve those physicians and biologists whose positions demand that they make the practical decisions."

Bioethics, then, has grown past its academic origins, if, indeed, its origins were academic. It has become, as Carl Elliott (2005, p. 380) puts it, "a self-contained, semiprofessional entity whose place in the bureaucratic structures that house it has become distinct – both from the traditional academic disciplines from which it emerged and from the clinical disciplines that it has sometimes aspired to resemble." As a result, it has become possible to work as a bioethicist without necessarily working as a professor, physician, or anything else. The bioethicist has come to garner "a certain amount of deference within the institution," dispensing ethical advice that many people working in the hospital feel they cannot ignore.

Judith Andre (1997, pp. 161–165), a philosopher by training but now engaged in bioethics, reflects upon bioethics precisely as a practice. By "practice," Andre means something like Alasdair MacIntyre's notion, developed in his book, *After Virtue*. As a practice or near-practice, Andre argues, bioethics should be evaluated not only for its scholarship, but more broadly for its practical impact. Does bioethics make the world a better place for the sick, and indeed for all of us? Andre argues that bioethics is not a subfield of philosophy because bioethics does not simply supply philosophical insights to health care. To be a practitioner of bioethics demands that one master a body of scholarly knowledge specific to bioethics, but also that one develop "interpersonal and institutional skills" that are necessary to communicate with people from a range of disciplines and walks of life. Andre's description is an apt one for what has become known as clinical bioethics. Indeed, interpersonal skills are probably more important than scholarly knowledge when it comes to

negotiating conflicts between family members. But Andre's comments only serve to confirm Jonsen's characterization of bioethics as a demi-discipline.

The term "bioethics" may have been born in the United States, but the practices of bioethics are engaged in throughout the world. Culture does, of course, shape discourse. Henk ten Have (2000, pp. 28–31) has noted that while some southern European counties have maintained a stronger emphasis on traditional medical ethics as "medical deontology," i.e., codes of conduct that are mixtures of ordinary moral rules, professional codes of conduct and rules of etiquette, northwestern European countries and the United States have emphasized problems in the doctor-patient relationship and moral issues created by the health care system, as well as public policy issues resulting from biomedical advances and research. Academic culture also shapes bioethical discourse. The different philosophical methodologies in the Anglo-American academy and in Continental Europe have also shaped the discourse differently, with Americans talking largely about justice, for example, while many in Europe focus on the notion of solidarity.

This diversity raises the important question of how different discourses and disciplines shape the universe of bioethics, and some scholars have been at work trying to analyze the situation. Edmund Pellegrino (1997, pp. 11–19) has described five models of how the disciplines that contribute to bioethics relate to one another. In the *traditional model*, ethics is taken as a philosophical discipline and bioethics is seen as a branch of philosophy. He sees this as closest to the "Georgetown model," as described by Warren Reich. The problem with this model, as Pellegrino points out, is that it is too narrowly conceived and risks missing the insights that the various other humanities can contribute to bioethics.

The *antiphilosophical model*, by contrast, reflects the antipathy of many both within philosophy and outside it to philosophical ethics. It tries to banish philosophy from bioethics altogether and replace it with one of the other disciplines. Pellegrino rightly worries that ethics without a philosophical basis will be reduced to "a species of moral gnosticism or intuitionism," or worse, "moral nihilism and relativism."

The *process model* is a procedural enterprise that "evades the conceptual issues." It emphasizes only the ways in which people go about trying to resolve moral issues. Thus it rejects identification of bioethics with any discipline and instead sees bioethics as a method for deliberation and decision-making. The process of collaborative deliberation is clearly necessary for bioethics, and Pellegrino recognizes this. But as he rightly points out, this is not enough. The purpose of moral reflection is "right and good conduct," and this will not necessarily come from process alone. The process itself must be subjected to critical analysis. Philosophy is the obvious discipline from which to conduct this critical analysis, but historical, psychological, and even scientific analysis may also play roles.

The *eclectic-syncretic model* corresponds in many ways to Potter's "Wisconsin model" of bioethics. Eclecticism recognizes merit in many different disciplines and moral viewpoints and selects from each what it sees as useful. Syncretism then tries to resolve the differences and fuse what it has chosen into a new system. This is, as we have seen, one of the hallmarks of interdisciplinarity. The general problem with this model, as Pellegrino recognizes, is that it robs each discipline of its specific contribution to the bioethical discourse. Ethics interacts with biology, with literature, with the law, with the social sciences, and with other disciplines to create the interdisciplinary bioethics. One prominent incarnation of the eclectic-syncretic model in today's medical humanities is the interaction of literature and ethics. Literature has much to contribute to our understanding of the human condition and of good and evil. It is especially important in its ability to evoke in us emotional responses to ethical demands. However, Pellegrino is right in pointing out that the rich moral content of literature does not confer any epistemological status on literature. As he says, "fictive characters are fictions." Literature can inspire us to be good; but literature can also inspire us to be bad. On its own literature cannot give the type of moral sanction and "complete account" of the moral life that is the very essence of moral philosophy.[1]

Finally, the *ecumenical model* allows philosophical ethics to retain its traditional identity, but also allows dialogue with literature, anthropology, history and evolutionary biology, all of which retain their own distinctive identities. All of these disciplines study the moral life, but each does so from a different perspective. These differences are precisely what make the bioethical dialogue so rich. The non-philosophical disciplines aptly describe the complexity, the context and the psycho-social aspects of moral behavior. Any ethical analysis must take these factors into account. But it is philosophy that has the power to examine "those conceptual elements and principles that transcend

detail.'' Thus, the ecumenical model makes bioethics closest to ethics traditionally considered, but enriches it by drawing in a broader range of human experience and reflection.

I think that Pellegrino's ecumenical model for bioethics is moving in the right direction. The medical humanities enrich bioethics greatly in the ecumenical model, yet philosophy retains a central position among the medical humanities, because it is the discipline that is rightly concerned with critical analysis of the moral claims and methodologies of other related disciplines. I would like to move even more, however, toward a model in which the philosophy of medicine has a central place in the metamedical multiverse. Thus, although the philosophy of medicine can be seen as a universe of discourse itself, it would also be the organizing force for the entire metamedical multiverse, including the universes beyond bioethics.

Philosophy of medicine

Henk ten Have (1997, pp. 105–106) has argued that the era in which bioethics was born and blossomed is also characterized by the virtual invisibility of the philosophy of medicine as a theoretical and practical endeavor. He attributes this invisibility to three interrelated phenomena. The first is the ''ethicalization'' of the philosophy of medicine. Instead of examining the range of philosophical issues raised by medicine, focus is increasingly put on ethical issues by people who ''have renamed themselves 'bioethicists'.'' The second is the ''technicalization'' of ethics. That is, bioethics is now seen as an autonomous discipline aimed at solving practical problems; it is no longer adequately characterized as moral philosophy. The third phenomenon is the anti-realism that is fostered by the stress of privatization, relativism and proceduralism. This is characteristic not only of bioethics, but more generally of post-modernism and in particular the social constructivism that is so prominent in science and technology studies. This is all in general agreement with the way I have characterized bioethics. I also concur with ten Have's (2000, p. 31) call for a ''broader philosophical framework'' for bioethics in order to connect the ''internal morality'' of medicine with the ''external morality'' of the social, cultural and religious traditions in which medicine is practiced.

Ten Have (1997, pp. 111–113) finds the origins of the philosophy of medicine in the nineteenth century and coming from a reinterpretation not only of medicine but also of philosophy. This was the time of the emergence of an organized medical profession, which could claim authority because of its scientific basis. But at the same time, philosophy also began looking to science for methodological and theoretical models for philosophical study itself. Thus, by the end of the 19th century, philosophers gave up the quest of constructing grand idealistic systems to explain medical reality. Instead, they shifted their attention to philosophical interpretation of the practices of medicine. Philosophy of medicine changed from a discipline offering alternative and competing theories of medicine, to a meta-discipline. Philosophy of medicine did not lose its connection with philosophy in general, however. The prominent Polish school of philosophy of medicine, for example, identified itself as Polish analytical philosophy and was particularly interested in clarification of language, logic, and epistemology. The Polish philosophers concerned themselves with analyzing very particular problems in great detail rather than constructing grand philosophical systems (ten Have, 1997, pp. 113–116).

Looking at the conceptual structures of philosophy of medicine over the past 100 years, ten Have (1997, pp. 116–119) identifies three major traditions. The *epistemological tradition* grew out of the characterization of medicine as a natural science and its increasing specialization. The theory and practice of medicine became radically separated, and the need for synthesis became a fundamental epistemological problem for the philosophy of medicine. Two epistemological strategies developed. One focused on organizing knowledge by focusing on rigorous methodology. The other focused not on methodologies that could produce objectivity and precision, but rather on appreciating the subjectivity of the knowing subject. The latter recognized that medicine was concerned more with acting than with knowing. The *anthropological tradition* predominated in Germany and the Netherlands from about 1930 until 1960. It emphasized the subjectivity not only of the knowing and acting physician, but also of the patient. Medicine is unique because it attends to the patient as a person. The *ethical tradition* has predominated since the 1960s.

All three of these perspectives should be included in contemporary philosophy of medicine. As ten Have (1997, pp. 119–120) recognizes, medical practice is embedded in society and culture, and so the essential nature of medical practice cannot be understood by the study of medical

science in isolation. This, ten Have claims, has two effects. First, it has changed the relationship between medicine and philosophy. Because medical practice is so directed by social influences and cultural values, it is no longer the province purely of physicians doing meta-reflection on their own practices. Second, medical practice cannot be understood without understanding the cultural values in which it exists. The question for philosophy of medicine is not simply one of what we know, but of what we want to do with our knowledge. For this, the epistemological, anthropological, and ethical traditions in the philosophy of medicine are all necessary.

How these perspectives are organized in the philosophy of medicine has become a matter of academic debate, however. This debate relates directly to the question of what is included in the philosophy of medicine universe – and whether it is expanding to be more inclusive, or contracting to be more exclusive.

The narrow view

Edmund Pellegrino represents a notable instance of a narrow view of the philosophy of medicine. He and philosopher David Thomasma proposed three ways in which philosophy and medicine interact (Pellegrino and Thomasma 1981, pp. 28–30). (1) Philosophy *and* medicine has to do with "mutual considerations by medicine and philosophy of problems common to both." For example, the mind-body problem set up by Descartes is an important problem for philosophers of mind, metaphysicians and epistemologists, but it is also an important concern for philosophers of medicine, who might have very different views of the problem itself stemming from particular concerns of medicine or medical ethics. In this model, philosophy and medicine address a common topic, but they remain independent disciplines in particular interests and methodologies. (2) Philosophy *in* medicine refers to the "application of the traditional tools of philosophy – critical reflection, dialectical reasoning, uncovering of value and purpose, or asking first-order questions – to some medically defined problem."[2] In other words, this model sees the contributions that the discipline of philosophy has made to critical thinking, framing questions, and other basic work of philosophy itself, and simply applies these methods to issues in medicine. (3) Philosophy *of* medicine is concerned specifically with the meaning of clinical medicine. Philosophy of medicine examines the conceptual foundations and ideologies of the clinical encounter of doctor and patient; thus, it really tries to provide a foundation for medical ethics. In a later paper, Pellegrino added a fourth category – medical philosophy – which is more literary than philosophical. This includes the informal or literary reflections of physicians on their clinical experience (Pellegrino, 1986, 1998). Essays of William Osler or short stories of William Carlos Williams would count as medical philosophy.

Philosophy of medicine, for Pellegrino, then, is restricted to the third model. The first model might take purely epistemological questions in medical research to be outside philosophy of medicine. On this account, such questions are more properly questions of philosophy of science or epistemology. These may have great importance for the practice of medicine, but they are not properly philosophy of medicine unless they directly contribute to the clinical encounter of doctor and patient. The second model is of interest only in the recognition that philosophy has provided methods for clear thinking; their application in medicine is important, but no more important than clear thinking in any facet of human life. The fourth model, medical philosophy, is more akin to the medical humanities in general. Philosophy of medicine proper, for Pellegrino (1998, p. 327), is concerned only with what is "peculiar to the human encounter with health, illness, disease, death, and the desire for prevention and healing." Philosophical concepts are studied only insofar as they relate to the human encounter with somatic or psychological well-being and dysfunction.

Arthur Caplan also sees philosophy of medicine in a narrow sense, albeit a very different one. In actually arguing that the philosophy of medicine does not exist, Caplan (1992) presents a narrow view. Referring to an early work of Edmund Pellegrino, Caplan distinguished philosophy *and* medicine from philosophy *in* medicine. The former includes medical ethics, bioethics, health policy, and medical aesthetics. An example of the latter is the study of professional codes by those in bioethics. But philosophy *of* medicine, for Caplan (1992, p. 69) is "the study of epistemological, metaphysical and methodological dimensions of medicine; therapeutic and experimental; diagnostic, therapeutic, and palliative." Caplan states that this is a stipulative definition. We can, of course, organize our pursuit of knowledge in any way we see fit, but the question is why we should accept this particular stipulation. Caplan's understanding of philosophy of medicine at first appears to be quite broad, but it really is not,

for its primary intent is apparently to exclude much of what others consider important to the philosophy of medicine. It is curious that it is limited to epistemological, metaphysical and methodological dimensions. Why should the philosophy of medicine not include aesthetic and ethical dimensions, when aesthetics and ethics are clearly part of the philosophical universe? Caplan seems to want to limit the philosophy of medicine to just those sorts of questions that the philosophy of science addresses. In fact, even in the argument against the existence of the philosophy of medicine, Caplan (1992, pp. 69–70) slips in this statement: "In short, the philosophy of medicine is a sub-discipline of the philosophy of science. Thus, its primary focus is epistemological not ethical, legal, aesthetic or historical."

A reasonable concern that both Caplan and Pellegrino have is in trying to limit the field so that it is not unnecessarily broad. While Pellegrino would narrow the focus to the clinical encounter, Caplan would narrow the focus to medical science. This latter strategy, however, narrows the focus too much. Certainly, part of the philosophy of medicine must concentrate on the issues that Caplan mentions. However, understanding aesthetics is as important to an analysis of plastic surgery as understanding epistemology is to an analysis of pathology and laboratory medicine. Both of these specialties are part of medicine. So, if Caplan's claim that epistemology should be a part of philosophy of medicine is correct, then aesthetics should also be a part of philosophy of medicine.

The broad view

A broader view of the philosophy of medicine is the one outlined by Schaffner and Engelhardt (1998). I take this broad view to be closer to what those who see themselves engaged in the philosophy of medicine are actually doing. On this account, philosophy of medicine is defined as "encompassing those issues in epistemology, axiology, logic, methodology and metaphysics generated by or related to medicine." The broadest conception includes medical ethics, although the authors recognize that this has become such a large topic that they do not specifically include it in their article. Elements of the philosophy of medicine that they do discuss include models of medicine, such as the narrow biomedical model or the broader biopsychosocial model of George Engel. Concepts of health and disease have been a "defining problem" for contemporary (and classical) philosophy of medicine. Whether these concepts are value-laden or not has been a source of ongoing debate. In addition, recent advances in molecular genetics challenge older views of normality and pathology. Investigations into the logic of diagnosis, prognosis and evaluation of therapies began in the 1950s and were extensively developed in the ensuing decades. Artificial intelligence programs led to computer-assisted diagnosis, and this became a source of rich philosophical discussion. Philosophical discussion has also focused on causation of disease and evaluation of therapies.

In fact, even those who hold a narrower view of the philosophy of medicine would endorse the importance of all the matters included in the broad view of the philosophy of medicine. Pellegrino's own work has touched on causality, logic and the mind-body relationship. These issues are taken to be important only insofar as they lay a foundation for medical practice and medical ethics, however. On the other hand, the broad view takes all these matters, including medical ethics itself, as part of the philosophy of medicine. Philosophy of medicine does contribute importantly to medical practice, but it goes beyond this in trying to understand theory as well.

Situating the discipline

As is the case with trying to understand the structure of the physical universe, the way one sees, or does not see, the philosophy of medicine in the metamedical multiverse depends to a great extent on how one interprets the data. Of course, how one interprets the data also is influenced by the way one sees the situation with regard to philosophy of medicine. The question how the philosophy of medicine is related to other fields was advanced by Arthur Caplan's paper (1992) arguing that the philosophy of medicine does not exist as a field. Even though there has been no diminution, and indeed a significant expansion, of scholarship in what appears to be philosophy of medicine in the nearly 15 years since Caplan's paper was published, the philosophy of medicine still struggles for recognition in the immense shadow being thrown by bioethics. In fact, Caplan has always recognized the importance of the philosophy of medicine, and part of the momentum that drove his paper was the recognition that the philosophy of medicine is sorely needed not only by bioethics, but also by the philosophy of science and by medicine itself.

Vic Velanovich (1994) argued that, even twelve years ago, philosophy of medicine had all the

characteristics of a developing field, even according to Caplan's criteria. The most problematic area, then and now, is the integration of the field into some "cognate areas of inquiry." Velanovich admitted that this was the most underdeveloped area, but drew on John Dewey to argue that the logical forms that govern a field of inquiry are developed as the inquiry itself proceeds (Velanovich, 1994, pp. 78–79). Thus, he admitted that Caplan's assessment of the state of the field may have been right at the time, but that the proper connections may emerge.

Twelve years later, philosophy of medicine activities are as robust as ever, yet as a field, it still seems to wander, not part of philosophy and not part of medicine, yet studied with great interest by members of both disciplines. Indeed, Caplan (2006) has recently argued that bioethics is an insufficient remedy for what ails contemporary medicine. He maintains that medicine needs to know what its methods are for dealing with bias and fraud so that it can resist the pressures put on it by "politics, money, ambition and greed." This is fundamentally an epistemological problem, and Caplan laments the fact that few physicians have any sophisticated knowledge of the philosophy of science or the philosophy of medicine. Philosophy of medicine may still not be a field, but Caplan obviously believes it is essential, at least in the narrow sense that he conceives it.

A related problem in defining philosophy of medicine as a field is figuring out exactly who is doing it. In a response to Caplan, Henrik Wulff (1992, pp. 79–81) distinguished several groups involved in matters pertaining to the philosophy of medicine. There are professional philosophers who have become interested in medical matters, physicians whose main interest has turned to philosophy, professional philosophers who have become very well versed in medicine, medical professionals who are also trained in philosophy, and medical professionals who devote themselves to medical practice. It is this last category, Wulff argues, that plays an important role in formulating problems for the philosophy of medicine. Wulff (1992, pp. 83–85) argues that Caplan fails to see the existence of the philosophy of medicine because he is looking at it from the perspective of a professional philosopher. This seems right, for philosophy has been reluctant to add the philosophy of medicine to its recognized list of sub-fields. However, Wulff (1992, p. 85) claims that philosophy of medicine is a "philosophical activity" that is "closely linked to the main trends of contemporary medical thinking." Because it "serves the same goal as the rest of medicine, philosophy of medicine should be seen as an "emerging (or reemerging) medical discipline." The trouble here is that the practice of medicine, a practical pursuit, is quite different from the practice of metamedicine, by definition a reflective pursuit. It seems much less likely that the medical profession will recognize the philosophy of medicine as a sub-discipline than philosophers will, for philosophy of medicine is much more like philosophy than medicine. To conclude, I would like to suggest a model of metamedicine that holds a broad view of the philosophy of medicine at its center.

Mapping the metamedical multiverse

Philosophy was traditionally regarded as the "queen of the sciences," standing in a unique place to establish foundations of knowledge and ultimate truths. Although those goals may no longer seem realistic, and although professional philosophy itself has sometimes wandered far from them, philosophy still occupies a central position inasmuch as it seeks the assumptions behind and essence of all human endeavors and seeks to integrate them. In this sense, philosophy of medicine might serve as the central metamedical discipline, reflecting upon and integrating the various disciplines that reflect on the science and art of medicine.

Van Leeuwen and Kimsma (1997, p. 100) rightly point out that medicine is both more than a science and less than a science. It is more than a science because it does not restrict itself to formulation of theories that hold under carefully circumscribed conditions; it is less than a science because it is confronted by the need to act even in the face of an uncertainty that is characteristic of medicine. Physicians bring to bear several different kinds of skills and knowledge on real problems, thus instituting a "medical discursive account of the patient's situation" (Van Leeuwen and Kimsma, 1997, p. 102). I believe that they are right in saying that philosophy, and especially European philosophy, provides crucial insights necessary to understand medical practice. If anything, providing such crucial insights is what makes philosophy of medicine distinctive, and in a sense, confirms Pellegrino's insistence that the clinical encounter is at the heart of the philosophy of medicine.

Specialization is obviously necessary, in metamedicine as well as in medicine, for understanding all

the aspects of such a complex practice is beyond any one individual or discipline. Indeed, Robert Neville (1974) argues that this ideal is impossible because the disciplines inhabit what he calls "different worlds." Each discipline selects elements as either relevant or irrelevant to the model of that particular discipline; the discipline then takes its own explanatory system to apply to the world as a whole and not just part of it. This allows the scientist, for example, to see science as the only discipline worthy of explaining the way the world is, with all other disciplines merely offering subjective opinions not worthy of being called knowledge. Nonetheless, Neville (1974, pp. 63–64) suggests that philosophy, which aims to cultivate the "richest possible experience" of the world, might serve the role of integration of knowledge by translating what those in the various disciplines are saying into an "integrating cosmology." Of course, these cosmologies would be only hypothetical, but they could be judged according to such values as comprehensiveness, ability to specify the terms of the various disciplines, applicability to the whole of experience, and internal consistency and coherence. This approach would be committed not so much to finding truth, but rather to providing a common language for various matters, theoretical and practical, arising from all the disciplines.

Thus, I would like to suggest an alternative model for our metamedical multiverse. The model that sees the medical humanities as a broad family containing the various disciplines is what, at first glance, seems obvious. Within the medical humanities, the various disciplines such as bioethics, philosophy, art, literature, and history of medicine, all inform one another to some extent, but remain worlds of their own, hence retaining their individual identities as disciplines. An alternative model, the bioethics model, tries to incorporate all the various disciplines within it to create a new interdisciplinary discipline called bioethics. The model I am suggesting is one inspired by Cardinal John Henry Newman. Newman (1996, p. 45) argues that all knowledge forms one whole that can be separated only by abstraction. All disciplines have a bearing on one another. For Newman, it belongs to philosophy as the "science of sciences" to comprehend "the bearings of one science on another, and the use of each to each, and the location and limitation and adjustment and due appreciation of them all, one with another." In a sense, it is philosophy in this sense (although not necessarily in the sense of professional philosophy as it is practiced today) that is the genuinely interdisciplinary field.

This model, somewhat analogous to Pellegrino's ecumenical model of bioethics, sees the philosophy of medicine as the core discipline, but not in the sense that bioethics tries to ingest all other disciplines. Rather, philosophy of medicine becomes the common language for all the medical humanities. I believe that taking philosophy of medicine, rather than bioethics, as central will benefit all the medical humanities by providing a broader foundation for analysis of this very complex realm of activity. Making the metaphysical, epistemological and aesthetic aspects of ethical decisions more prominent would provide for a much richer ethical discourse than is currently being fostered by the professionalization of bioethics. Bioethics as a practical endeavor is undoubtedly important, but it could be more.

This necessarily demands that philosophy of medicine be considered very broadly. It cannot just be a subset of the philosophy of science that looks at epistemological, metaphysical and methodological facets of medicine, as suggested by Caplan. Certainly these elements will be part of this broad philosophy of medicine, but they will not constitute the whole of it. Philosophy itself is a broad field – so broad, some might say, that it has ceased being one field. Nonetheless, I am suggesting a return to the roots of philosophy. That view is the one that gave rise to awarding the degree of doctor of philosophy to people who have studied in all sorts of fields, the humanities and the sciences. Thus, philosophy of medicine would offer reflection not only on the traditional philosophical problems inherent in medicine, but also on all of the medical sciences and humanities, and medical practice as well. I am suggesting neither a philosophical imperialism, nor that only professional philosophers will be capable of doing philosophy of medicine. I am only suggesting that philosophical thought about all the medical humanities and sciences offers the best hope at integrating a very broad field of scholarship and enabling at least some communication in a metamedical multiverse that is now characterized either by separate bubble universes that have much trouble seeing into other universes, or worse, by one big bioethical bubble.

Notes

1. This is not to say that bioethics must give a complete account of the moral life. Giving such an account is, however, just what moral philosophy tries to do. Martha Nussbaum (1990, pp. 138–143) has argued that traditional moral philosophy, or ethical theory, lacks the power to express all moral truths, and that literature is important in conveying some of these truths.

She thus distinguishes between ethical theory and moral philosophy, the latter being a more inclusive term, which would include both traditional ethical theory and literature (Nussbaum, 1990, p. 169, n. 2). I do not doubt the power of literature to convey truths in a way that abstract ethical theory cannot. However, it remains a fundamentally philosophical task to judge that what is being conveyed in the literature is indeed a moral truth.

2. It might seem that philosophical reflection on medicine constitutes "second order" reflection. But apparently the point is that in the philosophy *in* medicine model, first order philosophical questions are applied to medicine; it is only in the process of applying the first order questions that the reflection becomes "second order."

References

Andre, J., 1997, 'The Week of November Seventh: Bioethics as a Practice', in: R.A. Carson and C.R. Burns (eds.), *Philosophy of Medicine and Bioethics: A Twenty-Year Retrospective and Critical Appraisal*. Kluwer Academic Publishers: Dordrecht, pp. 153–172.

Callahan, D.: 1973, 'Bioethics as a Discipline', *Hastings Center Studies* 1, 66–73.

Caplan, A.L.: 1992, 'Does the Philosophy of Medicine Exist?', *Theoretical Medicine* 13, 67–77.

Caplan, A.L.: 2006, 'No Method, Thus Madness?', *Hastings Center Report* 36(2), 12–13.

Elliott, C.: 2005, 'The Soul of a New Machine: Bioethicists in the Bureaucracy', *Cambridge Quarterly of Healthcare Ethics* 14, 379–384.

Evans, H.M. and J. Macnaughton: 2004, 'Should Medical Humanities be a Multidisciplinary or an Interdisciplinary Study?', *Medical Humanities* 30, 1–4.

Fox, R.C. and J.P. Swazey: 2005, 'Examining American Bioethics: Its Problems and Prospects', *Cambridge Quarterly of Healthcare Ethics* 14, 361–373.

Greaves, D., 2001, 'The Nature and Role of Medical Humanities', in: M. Evans and I.G. Finlay (eds.), *Medical Humanities*. BMJ Books: London, pp. 13–22.

ten Have, H.A.M.J., 1997, 'From Synthesis and System to Morals and Procedure: The Development of Philosophy of Medicine', in: R.A. Carson and C.R. Burns (eds.), *Philosophy of Medicine and Bioethics: A Twenty-Year Retrospective and Critical Appraisal*. Kluwer Academic Publishers: Dordrecht, pp. 105–123.

ten Have, H.A.M.J., 2000, 'Bodies of Knowledge, Philosophical Anthropology, and Philosophy of Medicine', in: H.T. Engelhardt Jr. (ed.), *The Philosophy of Medicine: Framing the Field*. Kluwer Academic Publishers: Dordrecht, pp. 19–36.

Jonsen, A.R.: 1998, *The Birth of Bioethics*. New York: Oxford University Press.

Newman, J.H.: 1996 [1891], *The Idea of a University*, In: F.M. Turner (ed.), New Haven, Yale University Press.

Neville, R.: 1974, 'Specialties and Worlds', *Hastings Center Studies* 1, 53–64.

Nussbaum, M.C.: 1990, *Love's Knowledge: Essays on Philosophy and Literature*. New York: Oxford University Press.

Pellegrino, E.D. and D.C. Thomasma: 1981, *A Philosophical Basis of Medical Practice*. New York: Oxford University Press.

Pellegrino, E.D.: 1986, 'Philosophy *of* Medicine: Towards a Definition', *Journal of Medicine and Philosophy* 11, 9–16.

Pellegrino, E.D.: 1997, 'Bioethics as an Interdisciplinary Enterprise: Where Does Ethics Fit in the Mosaic of Disciplines?', in: R.A. Carson and C.R. Burns (eds.), *Philosophy of Medicine and Bioethics: A Twenty-Year Retrospective and Critical Appraisal*. Kluwer Academic Publishers: Dordrecht, pp. 1–23.

Pellegrino, E.D.: 1998, 'What the Philosophy *of* Medicine *Is*', *Theoretical Medicine and Bioethics* 19, 315–336.

Reich, W.T.: 1994, 'The Word "Bioethics": Its Birth and the Legacies of Those Who Shaped It', *Kennedy Institute of Ethics Journal* 4, 319–335.

Reich, W.T.: 1995, 'The Word "Bioethics": The Struggle Over Its Earliest Meanings', *Kennedy Institute of Ethics Journal* 5, 19–34.

Schaffner, K.F. and H.T. Engelhardt Jr.: 1998, Medicine, Philosophy of In: *Routledge Encyclopedia of Philosophy*. New York: Routledge.

Siegfried, T.: 2002, *Strange Matters: Undiscovered Ideas at the Frontiers of Space and Time*. New York: Berkley Books.

Van Leeuwen, E. and G.K. Kimsma: 1997, 'Philosophy of Medical Practice: A Discursive Approach', *Theoretical Medicine* 18, 99–112.

Velanovich, V.: 1994, 'Does the Philosophy of Medicine Exist? A Commentary on Caplan', *Theoretical Medicine* 15, 77–81.

Wulff, H.R.: 1992, 'Philosophy of Medicine – From a Medical Perspective', *Theoretical Medicine* 13, 79–85.

Part II

Philosophy of Technology:

Basic Concepts and Implications for Medicine

2

What Is Technology? Defining or Characterizing Technology

Why Bother with Definitions?

Many students, in my experience, especially in the natural sciences, are impatient with disputes about definitions. They are often called "merely semantic" and may seem hairsplitting. Indeed, they are semantic, in that they deal with meaning, but they are hardly trivial. Many apparently substantive disagreements really stem from the disputants having two different definitions of what is being discussed, say religion, but not being aware of it. Often people think that definitions are purely arbitrary; it means that effort need not be wasted on choosing among opposing or alternative definitions. This is itself based on one view of definition, but it is not the only one. We shall learn something about philosophy by seeing the different sorts of definitions that people have used and their connection to differing philosophical views.

Looking at the alternative definitions of technology shows something about the alternative kinds of definition and also about the characterization of technology. Even if one doesn't find a final definition on which everyone can agree, an investigation of the definition of technology shows us the range of things that can count as technology and some of the borderline cases where people differ on whether something should be counted as technology or not. Even an unsuccessful search for a best definition helps us to explore the layout of the area we are investigating.

As mentioned above, the major theorists of technology of the first two-thirds of the twentieth century believed that a universal, essential definition of technology could be given. A number of recent theorists, such as Don Ihde, Andrew Feenberg, and others, believe, in contrast, that there is not an essence or single defining characteristic of technology, and that searching for an essential definition is unproductive.

Guidelines for Definitions

Some general guidelines for definition are the following:

1. A definition should not be too broad or narrow. (That is, the definition should not include things we would not designate by the word we are defining, and the definition should not be so restricted as to exclude things that should fall under the term defined.)
2. A definition should not be circular. (For instance, we shouldn't define "technology" as "anything technological" and then define "technological" as "anything pertaining to technology.")
3. A definition should not use figurative language or metaphors.
4. A definition should not be solely negative but should be in positive terms. (A purely negative definition in most cases would not sufficiently limit the range of application of the term. A definition by contrast has to assume that the hearer knows the contrasting or opposite term.)

An example of defining technology in a too narrow manner is the common contemporary tendency to mean by "technology" solely computers and cell phones, leaving out all of machine technology, let alone other technology. A case of defining technology in a manner that may be too broad is B. F. Skinner's inclusion of all human activity in technology. Skinner understands human activity as being conditioned and self-conditioning. For Skinner conditioning is considered to be behavioral technology. A related move is the general inclusion of "psychological technology" as part of the motivational apparatus of technological activities, such as chanting in hunter-gatherer societies, or various political beliefs in industrial societies (propagated by propaganda, understood as a kind of technology by Ellul), thereby erasing the distinction between technology and culture by including *all* of culture within technology (see below on Jarvie).

Definitions of Technology

Three definitions or characterizations of technology are: (a) technology as hardware; (b) technology as rules; and (c) technology as system.

Technology as hardware

Probably the most obvious definition of technology is as tools and machines. Generally the imagery used to illustrate a brochure or flier on technology is that of things such as rockets, power plants, computers, and factories. The understanding of technology as tools or machines is concrete and easily graspable. It lies behind much discussion of technology even when not made explicit. (Lewis Mumford (1895–1990) made a distinction between tools and machines in which the user directly manipulates tools, while machines are more independent of the skill of the user.)

One problem for the definition of technology as tools or machines is cases where technology is claimed not to use either tools or machines. One such non-hardware technology is the behavioral technology of the psychologist B. F. Skinner (1904–90). If one considers verbal or interpersonal manipulation or direction of the behavior of another as technology then it appears we have technology without tools. Mumford claims that the earliest "machine" in human history was the organization of large numbers of people for manual labor in moving earth for dams or irrigation projects in the earliest civilizations, such as Egypt, ancient Sumer in Iraq, or ancient China. Mumford calls

this mass organized labor "the megamachine" (Mumford, 1966). Jacques Ellul considers patterns of rule-following behavior or "technique" to be the essence of technology. Thus, propaganda and sex manuals will be technology involving rules, and can, but need not always, involve use of tools or hardware.

Technology as rules

Ellul's "technique" mentioned above is a prime example of another definition of technology. This treats technology as rules rather than tools. "Software" versus "hardware" would be another way to characterize the difference in emphasis. Technology involves patterns of means–end relationships. The psychological technology of Skinner, the tool-less megamachine of Mumford, or the "techniques" of Ellul are not problems for this approach to technology. The sociologist Max Weber (1864–1920), with his emphasis on "rationalization," resembles Ellul on this, characterizing the rise of the West in terms of rule-governed systems, whether in science, law, or bureaucracy. Physical tools or machinery are not what is central; instead it is the means–end patterns systematically developed.

Technology as system

It is not clear that hardware outside of human context of use and understanding really functions as technology. Here are some examples:

1 An airplane (perhaps crashed or abandoned) sitting deserted in the rain forest will not function as technology. It might be treated as a religious object by members of a "cargo cult" in the Pacific. The cargo cults arose when US planes during the Second World War dropped huge amounts of goods on Pacific islands and cults awaited the return of the big "birds."
2 The Shah of Iran during the 1960s attempted to forcibly modernize the country. He used the oil wealth to import high technology such as jet planes and computers, but lacked sufficient numbers of operators and service personnel. It has been claimed that airplanes and mainframe computers sat outside, accumulating sand and dust or rusting, as housing for storage and the operating and repair staffs for them were not made available. The machinery did not *function* as technology.
3 Technological hardware not functioning as technology is not solely the province of indigenous societies or developing nations, but can also be

present in a milieu of high tech, urban sophisticates. Non-Western technology was displayed in an exhibit of "Primitive [*sic*] and Modern Art" at the Museum of Modern Art as purely aesthetic or artistic phenomena. Indigenous implements and twentieth-century Western abstract art objects were exhibited side by side to emphasize similarity of shape and design. The labels of the primitive implements often did not explain their use, only their place and date. (The use of these devices for cooking, navigation, and other purposes was not explained in the captions.) In some cases neither the museum visitors nor even the curators knew the technological function of the objects. Therefore, although the artifacts were simultaneously both technology and art for their original users, they were not technology, but solely art, for the curators and viewers of the museum exhibit.

These examples suggest that for an artifact or piece of hardware to be technology, it needs to be set in the context of people who use it, maintain it, and repair it. This gives rise to the notion of a **technological system** that includes hardware as well as the human skills and organization that are needed to operate and maintain it (see consensus definition below).

Technology as Applied Science

Much of *contemporary* technology is applied science. However, to *define* technology simply as **applied science** is misleading both historically and systematically. If one understands science in the sense of the combination of controlled experiment with mathematical laws of nature, then science is only some four hundred years old. Even the ancient Greeks who had mathematical descriptions of nature and observation did not have controlled experiment. The medieval Chinese had highly developed technology (see chapter 10) and a rich fund of observation and theory about nature, but had neither the notion of laws of nature nor controlled experiment. Technology in some form or other goes back to the stone tools of the earliest humans millions of years ago. Clearly, with this understanding of science and technology, through most of human history, technology was not applied science. Part of the issue is how broadly one defines science. If one means by science simply trial and error (as some pragmatists and generalizers of Popper's notion of conjecture and refutation have claimed; Campbell, 1974), then prehistoric technology could be treated as applied science. However, now the notion of science has

been tremendously broadened to include virtually all human learning, indeed all animal learning, if one holds a trial and error theory of learning. Perhaps this is an example of a definition of science that is too broad.

Even after the rise of early modern experimental science and the notion of scientific laws in the seventeenth century, and the development of the technology that contributed to the industrial revolution, most technological development did not arise from the direct application of the science of Galileo (1564–1642) and Newton (1642–1727). The inventors of the seventeenth and eighteenth centuries usually did not know the theories of mathematical physics of their day, but were tinkerers and practical people who found solutions to practical problems without using the science of their day. Even as late as Thomas Edison (1847–1931) we find a tremendously productive inventor in the field of electricity who did not know the electromagnetic theory of James Clerk Maxwell (1831–79) and his followers, but who produced far more inventions than those scientists who did know the most advanced electrical field theories. Edison initially even disparaged the need for a physicist as part of his First World War team, thinking one needed a physicist only to do complicated numerical computations, but that a physicist would have nothing much to contribute to technology. By this time Edison's view of the role of theory was getting somewhat dated.

Even in the contemporary situation, in which scientific training is essential for most technological invention, the notion of technology as applied science, if taken in too simple and straightforward a way, is misleading. Modern technology is pursued primarily by those with a scientific background and within the framework of modern science, but many of the specific inventions are products of chance or of trial and error, not a direct application of scientific theory to achieve a pre-assumed goal. Many chemical discoveries have been results of accidents. Safety glass was discovered when a chemical solution was spilled on a piece of glass laboratory apparatus, the glass was accidentally dropped, and it did not break. Penicillin was discovered when a bacterial culture was accidentally contaminated by a mold. Paper chromatography was discovered when a scientist accidentally spilled some chemical on a filter paper, and the chemical separated into two components as it seeped up the paper. The Post-it was discovered when a technologist, Art Fry, using little bookmarks in his hymnal, remembered a temporary glue that a colleague, Spencer Silver, had developed back in 1968 that was too weak to permanently stick two pieces of paper together. In 1977–9 3M began to market the invention, and by 1980 it was sold throughout the USA. Charles Goodyear's development of vulcanization of rubber

involved numerous trials and experiments, but one crucial event involved him accidentally leaving his treated "gum elastic" on a hot stove, and noticing that it charred like leather. He then experimented to find a lesser, but optimum, heat of exposure (Goodyear, 1855). Louis Pasteur (1822–95) famously said that chance favors the prepared mind. The development of these accidental discoveries made much use of the scientific knowledge of the people who made them. But the discoveries were hardly the straightforward application of scientific theory to a preset problem.

For these reasons, although technology involves knowledge, particularly know-how, a definition of technology that characterizes it simply as applied science is too narrow.

Systems Definition as a Consensus Definition of Technology

A number of writers have formulated a somewhat complex definition of technology to incorporate the notion of a technological system. The economist John Kenneth Galbraith (1908–2004) defined technology as "the systematic application of scientific or other knowledge to practical tasks" (Galbraith, 1967, chapter 2). Galbraith describes this as incorporating social organizations and value systems. Others have extended this definition to mention the organizational aspect of technology, characterizing technology as "any systematized practical knowledge, based on experimentation and/or scientific theory, which enhances the capacity of society to produce goods and services, and which is embodied in productive skills, organization and machinery" (Gendron, 1977, p. 23), or "the application of scientific or other knowledge to practical tasks by ordered systems that involve people and organizations, living things, and machines" (Pacey, 1983, p. 6). We can combine these definitions into "the application of scientific or other knowledge to practical tasks by ordered systems that involve people and organizations, productive skills, living things, and machines."

This consensus definition is sometimes characterized as the "**technological systems**" approach to technology. The technological system is the complex of hardware (possibly plants and animals), knowledge, inventors, operators, repair people, consumers, marketers, advertisers, government administrators, and others involved in a technology. The technological systems approach is more comprehensive than either the tools/hardware or the rules/software approach, as it encompasses both (Kline, 1985).

The tool approach to technology tends to make technology appear **neutral**. It is neither good nor bad. It can be used, misused, or refused. The hammer can be used to drive a nail or smash a skull. The tool user is outside of the tool (as in the case of carpenters' tools) and controls it. The systems approach to technology makes technology encompass the humans, whether consumers, workers, or others. The individual is not outside the system, but inside the system. When one includes advertising, propaganda, government administration, and all the rest, it is easier to see how the technological system can control the individual, rather than the other way round, as in the case of simple tools.

The notion (known as autonomous technology) that technology is out of human control and has a life of its own (see chapter 7) makes much more sense with technological systems than it does with tools. Technological systems that include advertising, propaganda, and government enforcement can persuade, seduce, or force users to accept them.

As noted above, not all students of technology wish to develop a definition or general characterization of technology. Some, particularly among the "postmodern" devotees of science and technology studies, claim not only that there is no "essence" of technology of the sort that mid-twentieth-century thinkers such as Martin Heidegger, Jacques Ellul and others claimed or sought, but that no general definition of technology is possible.

Despite the validity of the doubts of postmodern students of technology studies concerning an essence of technology, the "consensus definition" delineated above will help to keep the reader roughly focused on the kinds of things under discussion. For instance, the recent advocates of "actor-network theory" (see chapter 12) developed an approach to technology that has many affinities to the consensus definition in the technological systems approach. Advocates of the technological systems approach have recently begun to ally with or even fuse with the social construction of technology approach. Understanding technology as a network fits well with the European sociology of actor-network theory (see box 12.2). Thomas P. Hughes, the person who is perhaps the leading American historian of technological systems, has moved toward the social construction view, and combined it with his own approach (Bijker et al., 1987; Hughes, 2004).

PHILOSOPHY OF MEDICAL TECHNOLOGY

Sven Ove Hansson

1 INTRODUCTION

It has often been remarked that one of the foremost characteristics of modern medicine is its extensive use of technology. Medicine has always used technology, but since the late 19th century its reliance on technology has expanded dramatically. One of the many consequences of this is a change the location of the physician's activities. The use of special equipment made it necessary to move consultations from home visits to hospitals and physician's offices. As an example of this, the number of hospitals in the US increased from 200 to 4000 from 1873 to 1910 [Davis, 1981, p. 8].

Not surprisingly, various uses of technology figure prominently in discussions on medical ethics. However, few attempts have been made to give a comprehensive philosophical perspective on medical technology, and in the philosophy of technology medical applications are in fact marginal [Vos and Willems, 2000, p. 2].

Medicine and technology have much in common. Contrary to the natural sciences, neither of them is aimed at obtaining knowledge for its own sake. Both have an emphasis on techne rather than episteme, i.e. their goal is to find means of achieving practical results, to change the world rather than just to understand it [Hansson, 2007a; 2007b]. Medicine and technology also have a large and rapidly growing intersection, namely the use of technological methods to achieve the goals of healthcare. However, "[e]ven the most mechanical elements of medicine... are rarely, if ever, described as technology by its practitioners. Physicians are reluctant to see themselves as technicians or applied scientists" [Davis, 1981, p. 3].

The use of ever more complicated technology in hospitals has increased the role of engineers in healthcare. Engineers are responsible for the operation of essential diagnostic, therapeutic and palliative equipment. Due to the need for their expertise, some technological and engineering personnel are moving closer to the patient and assume more clinical roles in multidisciplinary healthcare teams [Deber and Leatt, 1986; Fielder, 1991; Wood, 2002]. Unfortunately, their role is often insufficiently understood by the public and by members of the more well-established healthcare professions. "Unlike other health professionals who have a firmly established role within the hospital system, clinical engineers often assume

Handbook of the Philosophy of Science. Volume 9: Philosophy of Technology and Engineering Sciences.
Volume editor: Anthonie Meijers. General editors: Dov M. Gabbay, Paul Thagard and John Woods.
© 2009 Elsevier BV. All rights reserved.

new and greater responsibilities without the needed authority or institutional support" [Saha and Saha, 1997, p. 189].

By investigating philosophical issues in medical technology, we can obtain a better understanding of clinical and biomedical engineering that are important branches of modern technology. Such studies will also help us to achieve a better understanding of the nature of medicine itself.

There are five major categories of medical or healthcare-related technology:

- *Diagnostic technology* identifies diseases and other conditions for treatment or palliation.
- *Therapeutic technology* is used in the treatment of diseases.
- *Enhancing technology* improves human functioning beyond what is needed to cure diseases.
- *Enabling technology* alleviates the impact of disease or a disability. This includes personalized equipment such as eyeglasses and artificial limbs but also universal technologies such as entrances that are accessible via wheelchair.
- *Preventive technology* reduces the risk or severity of accidents, toxic exposures, and other social and environmental mechanisms that give rise to disease or injury. This includes a wide variety of technologies, from sewage treatment plants to airbags.

Diagnostic, therapeutic, and enhancing technologies are integrated in healthcare. Enabling technology includes both technology that is part of healthcare, such as prosthetic technology, and technology that has little connection with healthcare. Preventive technology is usually not closely connected with healthcare, but in many cases, such as automobile safety, it makes extensive use of medical knowledge.

This chapter contains sections on diagnostic technology (Section 2), therapeutic technology (Section 3), enhancing technology (Section 4), and enabling technology (Section 5). Preventive technology is not treated here, but some aspects of it are discussed in *Risk and Safety in Technology* in part V of this handbook. The final Section 6 of this chapter is devoted to some issues that concern medical technology in general, namely how technology shifts responsibilities, what effects it has on the quality of care and human contact, and whether it gives rise to unsound and perhaps unnatural dependence on artificial devices.

2 DIAGNOSTIC TECHNOLOGY

Up to the 19$^{\text{th}}$ century, diagnosis was primarily an oral and visual process, unaided by instruments (the main exception being uroscopy). Physical diagnosis, often including measurements, was developed to a high degree of precision in the early 19$^{\text{th}}$ century [Davis, 1981, p. 183]. Around 1840 clinical laboratories were

introduced, offering an increasingly sophisticated repertoire of biochemical tests [Büttner, 2002]. In the 1880s and 1890s clinical photography rose to importance as a means of documentation. After Röntgen's discovery of X-rays in 1895 photography was overshadowed by X-ray diagnosis that had a deep impact on most clinical disciplines [Kröner, 2005]. Today, medical diagnosis is based on a combination of anamnesis (information obtained by interviewing the patient), physical examination of the patient, laboratory examination primarily of blood but also of other tissues and excretions, and imaging techniques including classical X-ray images, tomography and ultrasonography.

In recent years some types of diagnostic technology that were previously in the hands of physicians have been made available to the patients themselves. Asthmatics can use a peak flow meter to regulate their medication, and insulin-dependent diabetics can measure their blood-sugar levels and adapt the dosage. In particular the latter practice has had large impacts on therapy. With frequent measurements of blood sugar, blood sugar regulation has been made tighter, i.e. lower values can be kept without risking hypoglycaemia. This reduces long-term risks of blindness, neuropathy and atherosclerosis. It also makes it possible for diabetics to lead a less regular life, since they can adjust dosage to food intake and physical activity [Willems, 2000; Mol, 2000].

Technologically mediated progress in medical diagnosis gives rise to several important philosophical questions: How does increased diagnostic precision influence our concepts of disease? Is diagnostic precision motivated even when it does not lead to better therapy, or can it even have negative side effects? Can excesses in medical diagnosis give rise to social discrimination? The recent introduction of genetic technology in the clinical laboratory gives rise to further issues, in particular: Should we avoid collecting genetic information that may tell the patient more about herself than what she wants to know?

2.1 An excess of diagnoses?

Diagnosis is essential for treatment. Some of the most important contributions of technology to medicine have been diagnostic procedures that made it possible to offer patients more specific therapies and to commence therapy at an earlier stage of the disease. In some cases, the recognition of previously unknown preclinical signs of disease have made it possible to begin therapy before the patient suffers from the disease. Important examples of this are the use of mass radiography to discover early stages of tuberculosis and the use of sphygmomanometry to diagnose hypertension.

Not surprisingly, there are also cases when improved diagnosis has not been matched by corresponding developments in therapy so that, at least for a period of time, diagnosis has no effect on the patient's health. It has often been questioned whether diagnosis can have any value when it does not lead to a therapeutic intervention. In this discussion it is important to distinguish between two cases. The first case is *diagnostic information about a manifest disease.* Consider for

instance a patient with a back pain, who is referred to an X-ray exam. A possible outcome of the exam is the discovery of physiological changes in the spine that are not accessible to specific treatment and do not change the advice that the physician had already given the patient. Is such a diagnosis useless or perhaps even of negative value?

Experience from this particular diagnosis points in the opposite direction. Patients with back pain often want confirmation that their disease is real, and therefore appreciate knowledge about the physiological nature of the disease [Rhodes *et al.*, 1999]. Generally speaking, patients often want to know what disease they have. Furthermore, an exact diagnosis is in most cases required before the physician knows that it is useless to search for other, perhaps treatable, causes of the symptoms. Therefore, although not all diagnoses of manifest disease lead to improved treatment, careful diagnostication is usually an unavoidable component of responsible medical management of the patient's complaints.

The other, more problematic, case is that of a *diagnosis without a manifest disease*. Alvan Feinstein introduced the term *lanthanic disease* for diseases that can be detected by technological means, but are not experienced in any way by the patient [Feinstein, 1967; Hofmann, 2003]. Since the 19th century, life insurance companies have been a driving force behind the development of such diagnoses. They need methods to prognosticate a prospective customer's expected remaining length of life. Two technologies were shown in the early twentieth century to be efficient for this purpose, namely measurement of the person's blood pressure and her vital capacity (the maximal volume of exhaled air after a maximal inhalation). Physical standards based on sphygmomanometry and spirometry were used as health indicators in insurance medicine in the early twentieth century, but these diagnoses were not then matched by therapies [Davis, 1981, p. 185].

A modern example of a possibly problematic lanthanic diagnosis is osteoporosis at an early stage (also called osteopenia), as diagnosed through low bone mineral density (BMD, bone mass). This is an X-ray diagnosis (dual energy X-ray absorptiometry, DXA); the patient has no symptoms other than a somewhat increased risk of fractures. A study of women who received this diagnosis revealed that for many of them the bone scan had influenced their social lives. They perceived their bodies as fragile and therefore chose not to participate in a number of social activities. It is a widespread misconception that a person who suffers from osteoporosis should avoid physical activity in order to avoid fractures. In actual fact, the contrary is the case: physical activity is an important means of preventing an aggravation of osteoporosis [Magnus *et al.*, 1996; Dalsgaard Reventlow *et al.*, 2006]. Hence in this case, information about a technology-mediated diagnosis can be counterproductive in terms of medical prognosis. However, it is important to observe that this is not a necessary consequence of the use of this technology. Its effect will be positive if the physician who informs the patient of the diagnosis also manages to encourage her to increase instead of decreasing her physical activity, and to take other measures that contribute to halting the development of the disease, such as to stop smoking and reduce the intake of soft drinks.

2.2 Diagnosis as a source of social discrimination

New diagnoses often have impact on our concepts of disease and health, and they can also influence the way in which we conceive our bodies [Vos and Willems, 2000]. Hence, the exact measurement of physiological functions has led to new concepts of normality and abnormality, such as the notion of hypertension [Davis, 1981, p. 5]. New diagnoses can also be used to classify persons in new ways. Such classifications can have negative social effects for the persons to whom they are applied; in particular they can be used to discriminate against the persons so classified.

Discrimination means that certain persons receive a worse treatment, or less of some advantage, than others, without sufficient justification to select them for such inferior treatment. The most discussed types of discrimination are those that affect women, ethnic, religious, and sexual minorities, and people with certain handicaps and medical conditions. In some cases a diagnosis alone, i.e. a diagnosis without an accompanying actual condition, can have a discriminating effect [Hansson, 2005].

The clearest evidence of such discrimination can be found in the insurance sector. Insurance companies have a right to collect medical information about their customers. They also have economic incentives to use such information to the customers' disadvantage. Hence, patients with hereditary hemochromatosis have found themselves excluded from insurance although they complied with therapeutic phlebotomy and therefore had no increased risk of disease or death. (Some relatives of patients with this diagnosis have avoided such discriminatory treatment by not having themselves tested but instead donating blood as often as phlebotomy is recommended for patients with the disease [Barash, 2000]). Afro-Americans who are carriers of the sickle-cell trait have been discriminated against by life insurers, although their condition does not give rise to an increased risk of death [Bowman, 2000].

It should be emphasized, however, that the extent to which insurance companies have incentives to discriminate customers with certain diagnoses depends on the politically chosen construction of the insurance system. Hence, the American insurance industry uses such information to reject applications for health insurance policies and to refuse payment for the treatment of illnesses [Alper and Beckwith, 1988; Anderlik and Rothstein, 2001]. The prevalence of this practice depends on the fragmentary nature of American health insurance [Wolf, 1995]. Most European countries have more developed health insurance programmes that cover everyone and have the same premium for all persons on the same income level. In such systems there is no incentive for health insurers to collect prognostic medical information about their customers. On the other hand, the system for life insurance seems to be more or less the same in all countries, and gives rise to such an incentive.

Another situation where discrimination can be based on a diagnosis is the recruitment of personnel. Employers can require medical information about prospective employees. A well-known example concerns the sickle cell gene. The U.S. Air

Force barred Afro-Americans with the sickle-cell trait from becoming pilots due to an erroneous belief that they were prone to illness at high altitudes [Dolgin, 2001]. In later years worries have been expressed that genetic information can be used by employers to discover predispositions to certain diseases, recessive genes for inherited diseases, or (hypothetically) various psychological characteristics [Brady, 1995; Silvers and Stein, 2002; Persson and Hansson, 2003]. However, it should be emphasized that the use of diagnostic technology for such purposes is within social control. Several countries have passed laws that regulate what information an employer may acquire about a prospective employee.

One of the best-known examples of maltreatment based on a mere diagnosis is the social discrimination of recessive carriers of the sickle-cell gene in the Greek village Orchemenos. Since the gene was unusually common in this village, all inhabitants were offered testing. The purpose was to make it possible for carriers of the gene to avoid marrying other carriers. However, this strategy failed, and instead testing led to stigmatization of the carriers. Non-carriers chose to only marry other non-carriers, and carriers were left to marrying each other [Moore, 2000]. Another example is the Ashkenazi Jews. This group has a long history of volunteering for genetic research, and therefore a disproportionate number of genetic alterations have been shown among them. This has given rise to a widespread though mistaken view that they are more prone to genetic disorders than others, and they have on occasions been discriminated for that reason [Dolgin, 2001].

2.3 Genetic diagnoses

In recent debates about discrimination it has usually been taken for granted that genetic information is more sensitive than most non-genetic information. The use of genetic information is also much less accepted. While it seems to be fairly accepted that a person who has a manifest illness with a bad prognosis is denied a life insurance, rejections based on genetic tests have been vehemently protested against. The view that genetic information requires more protection to ensure privacy than most other forms of medical information has been called genetic exceptionalism [Green and Botkin, 2003].

Genetic exceptionalism is an example of a general tendency that is also seen in many social and ethical debates on biotechnology: The application of technology to a genetic material is conceived as particularly sensitive and is sometimes seen as ethically problematic in itself.

More concretely, three major differences between genetic and non-genetic information have been invoked to defend genetic exceptionalism. First, genetic information is said to give more precise information about the likelihood of future disease than what is obtainable from non-genetic tests. Secondly, genetic tests provide information not only about the tested individual but to some extent also about her relatives. Thirdly, genetic information is said to reveal fundamental and immutable characteristics of the individual [Alper and Beckwith, 1988].

As one example of the first argument (the predictive power of genetic tests), Roche and Annas [2001] claim that DNA-sequence data differs from other types of medical data in providing information not only about a patient's current health status but also about her future health risks. According to these authors, genetic information is in this sense analogous to a coded "future diary". This, however, is a severely misleading statement. Although information about single-gene diseases may have a high predictive power, most health-related genetic information refers to diseases with a complex etiology involving several genes and several environmental factors. In such, more typical cases the predictive power of genetic tests is far from impressive. There are also several examples of non-genetic diagnostic technologies with a high degree of predictive power. Two practically important examples are sphygmomanometry and tests for fecal occult blood. They both have great value in detecting diseases (hypertension respectively colon cancer) in their early stages before the patient is aware of it.

Concerning the second argument, it is certainly true that family members can be affected by results from genetic tests. However, the same applies to non-genetic tests for infectious diseases (not least sexual partners in the case of sexually transmitted diseases). An interesting comparison can be made been made between Huntington's disease and HIV in this respect. Huntington's disease is a rare genetic neurological disease that usually does not give rise to noticeable symptoms until the patient is in her thirties or forties. Having the abnormal Huntington gene is similar to being HIV-positive in at least two important respects: One may remain healthy for a number of years before the onset of the disease. Furthermore, both conditions are frequently transmitted to offspring [Gin, 1997].

Finally, concerning the third argument, genetic information is believed to reveal who the person "really is". This view of personhood has been called "genetic essentialism" [Alper and Beckwith, 1988]. According to that view, genetic information is more intimately related to a person's true nature than other sorts of information about the person. As Launis [2000] has argued convincingly, genetic essentialism is based on the highly controversial metaphysical presumption that there is such a thing as a person's core nature, or essential identity. Furthermore, the available empirical evidence shows that we are constituted by a combination of genetics and environment, not by genetics alone.

However, it is possible that the technological availability of genetic information will lead to more emphasis on genetic, inherited aspects not only of health but also of human personality. In this way, technologically mediated knowledge might have impact on how we view each other as persons: It might lead to a focus on inherited, unchangeable traits rather than on the social influence on personality.

On the other hand, other technologies are also developing that may have an opposite effect. Proteomics, and information about the expression rather than the presence of a gene, may become more predictive than genetic sequencing. Biochemical tests can be developed that reveal environmental influences on the person. The development of future diagnostic technologies will in all probability provide us with tools that reveal both the genetic and the environmental influences

on our bodies and our personalities. It is not possible to predict in what way these developments will influence our views on human beings, but the philosophical impact may be substantial.

3 THERAPEUTIC TECHNOLOGY

Therapy, the remediation or treatment of a health problem, is of course at the centre of medicine (although the prevention of disease or accidents is no less important). Therapy has always involved technological procedures; fairly advanced surgery such as trepanation was performed in Neolithic times.

3.1 Therapeutic knowledge and knowledge of side-effects

Today it is taken for granted, at least in academic medicine, that therapy should be based on scientific knowledge. However, the connection between therapy and science is much more recent than that between therapy and technology. In Hippocratic medicine that dominated medicine for more then two millennia, the most common therapies were bloodletting, purging, and emetics, all of which were positively harmful to the patients. Although medicine has been taught in universities since the late thirteenth century, its practice was based on Hippocratic teachings. Important advances in understanding of human biology were made, such as Harvey's discovery of the circulation of blood, but they led to no therapeutic advances [Wootton, 2006]. It was not until the nineteenth century that professors of medicine strove to make their discipline one of the sciences. Two major approaches were taken to achieve this. One was to make medical therapy essentially a branch of the natural sciences. By studies in the laboratory, diseased organs and tissues could be classified and causes of disease could be revealed. Claude Bernard was a leading proponent of this approach to the scientification of medicine. The other approach was treatment experiments, i.e. what we today call clinical trials. In the nineteenth century the first pioneers of clinical research began to evaluate the effectiveness of therapeutic methods through statistical comparisons of groups of patients who had received different treatments [Booth, 1993; Wilkinson, 1993]. Originally, the two approaches to scientific medicine were seen as competitors. Today it is generally recognized that laboratory research is as necessary to develop new therapies as is clinical research to validate, evaluate, and calibrate them.

Hence, the crucial source of therapeutic knowledge is the clinical trial. In a clinical trial, groups of patients with the same disease receive different treatment, and statistical analysis is performed to determine both the therapeutic effects and the side effects in the different groups. In this way, the therapy with the best balance between therapeutic chances and (risks) of side effects can be identified. The ethical defensibility of clinical trials is far from self-evident. The consensus view is that a clinical trial is only acceptable if there is genuine uncertainty about which of the tested treatments is best, and informed consent has been obtained from all the subjects [Hansson, 2006].

Although clinical trials were proposed in the early nineteenth century, they were rare until after World War II. Today, a large part of the published medical research is reports from clinical trials. Since the 1990s, the use of information from clinical trials for clinical decision-making has been facilitated by the development of systematic procedures for evaluating clinical research (evidence-based medicine, EBM) [Evidence-Based Medicine Working Group, 1992].

The vast majority of clinical trials concerns pharmacological treatment. A major reason for this is that new drugs are not allowed unless they have been shown in clinical trials to be therapeutically useful in comparison to previously available therapy. Government control of medical devices is less extensive than for pharmaceutical products. In particular, there is no general system for premarketing testing similar to that for drugs [Altenstetter, 2003]. As a consequence of this, much less clinical research is performed on the therapeutic use of technical devices than on the therapeutic use of drugs.

Not surprisingly, mechanical and other technological devices can fail in unforeseen ways, just like drugs. There is a long historical list of such failures. The majority of these did not give rise to severe injuries. But there have also been cases when technological failures had fatal outcomes. One of the best-known cases is the Bjork-Shiley heart valve, in which case regulators and industry seem to have been too slow in taking actions to prevent continued implantation of a defective product. The decision to withdraw the product came unnecessarily late according to critics. The decision was not made by the regulators but voluntarily by the company [Fielder, 1991].

It is important to relate the producer's responsibility for the functioning of a device to the actual clinical settings in which it will be used. One critic complained that "most medical device designers appear to have envisioned the controlled, delicate, and precise choreography of a surgical team, not the frantic activity of the emergency room or a 'code-blue' call. Consequently, many devices are not as rugged and easy to use as they could be" (Houston, quoted in [Saha *et al.*, 1985]).

However, this situation may change. One observer of the system described the current situation as follows: "The long-lasting honeymoon between the industry and European healthcare regulators seems to have ended. For healthcare payers and purchasers the case is clear: medical technology is a cost-driving force. Thus, medical devices and the medical device industry have come under increasing scrutiny and regulation" [Altenstetter, 2003]. A possible outcome of such increased scrutiny could be that more clinical trials are undertaken in order to determine the functionality of therapeutic technology.

3.2 Therapy vs letting die

Discussions on death have a central role in medical ethics, and they have often been connected to critique of technology. Some critics see the "modern" death in a technologically equipped hospital as "unnatural", whereas they regard "natural" death without modern medical technology as more dignified. This is a highly ro-

manticized view. "Natural" death is often an extremely painful process, whereas modern technology can to some extent relieve the dying person of pain and distress [Barnard and Sandelowski, 2001].

Many critics also underestimate the quality of life that is obtainable with life-sustaining technology. Hence, it is often believed that a life with a ventilator could not be worth living. In actual fact, long-time use of a ventilator is perfectly compatible with a good quality of life [Bach and Barnett, 1994].

However, even after the exaggerations have been removed, difficult ethical problems remain in the use of medical technology on severely ill patients. Just as there are occasions when permanent use of a ventilator can help a patient to a meaningful life, there are also occasions when the use of a ventilator will keep alive the body of a person whose brain does not function any more. The issue of futility, and what technological means are justified in the treatment of a severely ill person, is mainly a medical issue. The crucial criteria are the patient's condition and prognosis, in particular her level of consciousness, and her own preferences as far as they can be known. However, there are also some technological aspects to this question.

One such issue is the distinction between act and omission, and correspondingly between causing someone's death actively and causing it by refraining from doing something (e.g. refraining from a therapeutic action that is considered to be futile). This distinction has crucial role in the debate on euthanasia, but it is nevertheless far from clear [Hansson, 2008]. Hence, a physician who withdraws a respirator from a terminal patient with no hope of recovery is often seen as (passively) permitting death to occur through natural causes. In contrast, a well-meaning friend or relative who disconnects the respirator would run much greater risk of being accused of killing the patient. It seems as if the distinction between killing and letting die depends on social conventions and role norms [Winkler, 1988].

The withdrawal of nutrition from a terminally ill patient seems to be particularly problematic. It is an important part of medical and nursing tradition that patients should be given basic care and comfort even when the progress of the disease cannot be prevented or delayed. This includes the provision of food and fluid. Therefore, some maintain that the terminally ill should be provided with nutrition and water, even if this has to be done by technological means rather than by feeding them and giving them to drink. Others are unwilling to extend the requirement to provide nutrition and hydration to cases when this can only be done with a nasogastric tube or intravenously [Winkler, 1988, p. 165].

The continued use of new advanced devices on terminally ill persons has sometimes been questioned. This applies in particular to left ventricular assist devices (LVAD) and total artificial hearts (TAH). Although originally intended as bridging devices, LVADs have been used as destination therapy with good results. Total artificial hearts are at the time of writing still essentially an experimental therapy. Consider a case when an LVAD has been implanted as a bridging device, but circumstances have changed so that transplantation is no longer an option. It could then be claimed that since the device is no longer medically indicated,

it can be turned off or removed. However, both of these actions are expected to hasten the death of the patient [Bramstedt and Wanger, 2001]. Switching off the device under such circumstances would be contrary to generally accepted ethical principles. The same problem arises, perhaps in more drastic form, for total artificial hearts. Katrina Bramstedt has claimed that "the fact that a TAH (or any other implant or assist device) is functioning without flaw is of no relevance to the futility discourse. What is relevant to these discussions is whether the 'perfectly' functioning device is serving the goals of medicine and the best interests of the patient. Just as with a ventilator, a TAH can be functioning 'perfectly', yet be ethically inappropriate." Furthermore, she says that "[a]s with implantable defibrillators, inactivation of a TAH is a simple procedure not involving surgery, and this inactivation should not be seen as ethically separate from the withdrawal of other life support measures such as dialysis or ventilation" [Bramstedt, 2003]. A contrary view was expressed by Robert Veatch [2003], who claims that Bramstedt "appears to be endorsing unilateral actions by physicians that will directly cause the death of their patients and do so against the will of the patient or surrogate. That should be called 'murder'." According to Veatch, "[t]hrowing a switch that stops a TAH is more like injecting a drug that paralyzes the heart muscle or like excising the SA node. Either of these would be considered direct, active killing. How can it be that turning off the heart is any different?" Whereas other authors have emphasized the similarity between turning off an artificial heart and discontinuing other life-prolonging treatment [Miles *et al.*, 1988], Veatch emphasizes the difference. It could be argued in favour of his view that a patient who has received an artificial heart will regard it as her own, and thus not as a device that somebody else has a right to stop.

Future technological developments may provide us with other types of life-sustaining devices that give rise to essentially the same type of questions as the artificial heart. This would apply, for instance, to an artificial lung or kidney. A somewhat different type of end-of-life issue would arise from a brain implant that is not necessary for life but necessary to support consciousness. If the quality of the achieved consciousness deteriorates, arguments could be made in favour of turning off such an implant. This would, however, be a highly problematic standpoint for same reason that turning off a life-sustaining artificial organ is problematic.

3.3 Subcultures that resist therapy

Medical technology has effects not only on individuals but also on social groups and on society as a whole. Radical improvements in treatment will change the situation of disabled subcommunities in our societies. Perhaps surprisingly, therapeutic improvements are not always received positively in these subcommunities. The "fat is beautiful" movement denies that obesity is a disease requiring treatment and medical attention. Segments of the dwarf community have reacted against the introduction of therapies against their condition, seeing this as a threat to the future existence of their way of life and their organizations [Berreby, 1996].

By far the strongest such counter-reaction is the criticism from the Deaf World of cochlear implant surgery in prelingually deaf children [Crouch, 1997; Lane and Bahan, 1998].

The criticism of cochlear implantation is associated with a positive view of deafness. The Danish Deaf Association has stated that "deaf children are not sick or weak children, but normal Danish children, who just happen to use another language" (quoted in [Nunes, 2001]). Members of the Deaf World reject the idea that they have an impairment or disability. Instead, they view themselves as a minority culture with its own language, customs, attitudes, knowledge, and values. The use of cochlear implants will lead to a drastic decline in the population of this minority culture. Deaf activist have often referred to the ethical principle that minority cultures should be preserved. They claim that large-scale implantation of children conflicts with the right of the Deaf language and cultural minority to exist and flourish. The term "genocide" has sometimes been used to describe that prospect [Lane and Bahan, 1998].

This claim has given rise to an interesting discussion about the definition of a minority culture and whether cultures have intrinsic value [Levy, 2002]. Critics have pointed out the problematic nature of arguments that give precedence to the preservation of a culture over the interests of individual children. Some have noted that it is difficult to draw the line if cochlear implants are disallowed for this reason. If cochlear implants are unethical, then how should we judge the rubella vaccine [Balkany, 1996]?

From the viewpoint of mainstream medical ethics the interests of a subculture that needs to recruit new members could hardly prevail over the physician's responsibility towards the individual patient. Nevertheless, there are important lessons to be drawn from this debate. In particular, it shows that the ethical discussion on medical technology must take into account the social and cultural notions of disease.

4 ENHANCEMENT TECHNOLOGY

Technological devices such as implants can be constructed not only to cure disease and restore human functioning to normal levels, but also to improve human functioning to levels above the normal. The philosophy of medical technology therefore has to deal with issues of normality and disease and with the admissibility of human enhancement. If it becomes possible to improve a healthy person's physical strength or her memory to levels above her natural endowment, to what extent is it advisable to do so?

4.1 Enhancement and the limits of normality

Much of the recent debate on enhancement has referred to genetic enhancement, which only few writers defend [Resnik, 2000]. In this area, the enhancement discussion is anticipatory since no enhancing genetic technology is currently available.

However, there are at least two branches of medicine that already deal with enhancement in everyday clinical decisions, namely cosmetic surgery and neuropharmacology. Many types of cosmetic surgery, including breast implants, have been criticized for not complying with the aims of medicine, since they do not treat a disease or malfunction [Jacobson, 1998; Miller *et al.*, 2000]. Several drugs developed to treat diseases of the nervous system also have the ability to improve normal functioning. Hence, drugs developed for the treatment of narcolepsy are already in use in armed forces as wakefulness drugs. Drugs against depression are used for mood elevation by persons with no psychiatric diagnosis, and drugs against erectile dysfunction are used for pleasure [Wolpe, 2000]. Drugs developed to prevent cognitive deterioration in Alzheimer's disease seem to be capable of improving cognitive functioning in the healthy.

In addition to enhancement of capabilities that we already have, it is also possible to develop entirely new functions for the human body. Currently, microchip devices are implanted in animals for identification purposes. It is technically possible to implant similar devices into humans. One use of such chips would be to let airplane passengers travel without a ticket or identity document; instead they would be scanned. A more sophisticated read-write chip could carry a person's medical history or her criminal record. An implanted radio transmitter can be used to track a person [Ramesh, 1997]. A related prospect is that of implanting a device in the body that continuously monitors levels of substances in the bloodstream, and adjusts drug release accordingly [Wood *et al.*, 2003].

Some authors are against virtually all forms of enhancement since it transcends the traditional task of medicine that is to treat and prevent diseases, not to improve humanity generally. "[T]he goals of medicine concern not all human suffering, but only that suffering connected with a malady" [Miller *et al.*, 2000]. There are at least two problems with this standpoint. First, the distinction between disease and health or normality is not as clear as it may first seem. Disease is not a biologically well-defined concept but one that depends largely on social values. Some conditions previously regarded as diseases are now regarded as normal states of the mind or body. Other conditions that were previously regarded as variations within normality are now regarded as diseases. Homosexuality is an example of the former, attention-deficit hyperactivity disorder (ADHD) an example of the latter.

Secondly, it is easy to show with examples that our intuitions about whether treatment should be offered for a condition are strongly influenced by other factors than whether or not that condition is classifiable as a disease. One well-known example is the treatment of short stature. Both public and private insurers have chosen to pay for growth hormone treatment only if the child has some diagnosable growth hormone deficiency, not otherwise regardless of how short it is projected to be [Verweij and Kortmann, 1997]. As was noted by Norman Daniels [2000], this criterion for treatment is difficult to defend from an ethical point of view. If one person is short "just" because of her genotype and another due to some identified dysfunction, this does not mean that the first person suffers less or needs treatment less. Clearly, neither of them is short through a choice or fault of her own.

(In practice, however, we have been saved from ethical predicaments of growth hormone therapy by studies showing that this treatment does not affect the final, adult height of children who have a normal endogenous production of the hormone [Murray, 2002].)

Presbyopia is a normal feature of aging, and should therefore not be regarded as a disease. Nevertheless, we do not hesitate to treat this condition (mostly with eyeglasses). Hopefully, no one would try to prevent ophthalmologists from treating this or other age-related conditions of the eye. Now suppose that a remedy becomes available for age-related cognitive decline. It is a good guess that — perhaps after some initial hesitation — our attitude to such a treatment would be the same as to presbyopia. (Or would anyone say: "Just let grandmother become confused. It is not a disease, so although there is a treatment she should not take it. Treatments are only for diseases.")

We already endorse improvements of the immune system (vaccinations). Other ways to improve the body's resistance against disease would probably find acceptance relatively easily. There are also situations in which improved cognitive function would be seen by most of us as an advantage, such as improved driving ability and improved ability of surgeons to operate [Whitehouse *et al.*, 1997].

It is also interesting to compare our views on improvements of the teeth and of the skin. In the middle of the 19th century it was normal for nearly all an adult's teeth to display signs of decay. At that time, the type of dental work that is now routine would have been seen as remarkable and perhaps even as ethically doubtful. Today, it is about as difficult to provide old people with skin that looks youthful as it was then to make their teeth look youthful. How will we react if future developments make wrinkled skin as avoidable as discoloured tooth stubs are today?

These examples show that the disease/normality limit does not tell us what treatments are acceptable. However, there may still be other arguments against enhancement, arguments that do not depend on the distinction between disease and normality. One obvious such argument is that enhancements may have serious side effects. Hence, we can expect genetic enhancement to have unknown negative effects [Goering, 2000]. In one experiment, mice that were genetically engineered to improve their performance on learning tasks turned out to have greater sensitivity to pain [Wei *et al.*, 2001]. Perhaps a method to improve memory will have psychological side effects since it prevents us from forgetting things we cannot bear to think about. "Who needs to remember the hours waiting in the Department of Motor Vehicles staring at the ceiling tiles, or to recall the transient amnesia following a personal trauma" [Wolpe, 2000]? Other side effects may follow from other types of enhancement. However, although this type of argument can be used against many methods of enhancement, it is not a decisive argument against enhancement as such.

At the bottom line, the enhancement issue concerns what kinds of human beings there should be. Should future people be stronger and more intelligent than we are? A common, often religiously motivated view is that human nature has been given

to us and should not be changed. Others see considerable scope for improvement of the human race. In one of the few scholarly papers devoted to the issue, James Hudson maintains that to the extent that we can influence the innate natures of future people, we should make them intelligent and probably without a sexual drive or "*any drive...* other than a drive to *rational thought and action* in general" [Hudson, 2000]. Needless to say, this is a controversial standpoint.

The issue what kind(s) of persons there should be is among the most difficult ones to deal with rationally in moral philosophy. The very basis for the discussion is insecure. What criteria should we use? Should we judge future persons by our own criteria, or by the criteria that we predict (and partly determine) them to have? (Population ethics that deals with how many persons there should be has similar difficulties.) Possibly, the best way to tackle issues of enhancement is to deal with them incrementally, judging each individual case on the basis of our current values without even trying to take future values into account. However, such incrementalism needs to be informed by a discussion about possible long-term developments. The following words of warning are worth taking into account:

> Whereas one can make the case that future generations should have the right to decide by themselves about their fate, it should be prevented that we enter a slippery slope towards ever greater manipulation of the human body, without medical necessity, and do so without having fully considered the consequences. [Altmann, 2001]

4.2 Making man-machines

Microprobes implanted into nervous tissue can create interfaces for communication between a patient's nervous system and devices that replace or supplement a malfunctioning organ. Currently the most important of these neural interface implants are cochlear implants (see above, Section 3.3). Brain implants are also used for bladder control and for blocking tremors for instance in Parkinson's disease. There are several other promising applications, including the control of epileptic seizures [Pereira *et al.*, 2007]. Experiments have been performed with chips implanted in the brain or a peripheral nerve in order to control a wheelchair or other compensatory technology, or a prosthetic device such as a prosthetic hand [Warwick *et al.*, 2003; 2007]. Research is being conducted on prosthetic vision for the blind, based on essentially the same principles as cochlear implants, namely that stimuli from technological sensors are relayed to the nervous system via a nerve-implant interface. Two major alternatives are being investigated for the placement of this interface, namely retinal chips and chips implanted in the visual cortex of the brain. Prosthetic vision is currently primarily developed in animal models, but preliminary testing on human volunteers has taken place [Bertschinger *et al.*, 2008; http://www.bostonretinalimplant.org].

If efficient implantable brain chips become available, then they can be used for various forms of enhancement. It has been speculated that military applications

can come first, with the purpose of producing soldiers with enhanced abilities [Maguire and McGee, 1999; Altmann, 2001]. Some computer visionaries dream of a future in which many or all humans have implantable computer chips that connect them to sensors, assist their memory, and provide them with a variety of capacities. The "cyborgs", cybernetic organisms, of science fiction that are mixtures of man and machine would then become reality [Behling, 2005]. Some authors have hailed this as a positive development, since cyborgs can become better than men [Haraway, 1991].

It has also been argued that such neural implants could in the future be used to scan, upload and transfer (the contents) of a mind. Computer-brain connections will then allow electronic communications with other similarly connected individuals in a way that may require that we radically reassess the boundaries between self and society. However, this is even more speculative than the idea of a cyborg. We do not know whether or not complex sensory impressions, feelings and thoughts, can be communicated in either direction through an implant [White, 1999].

5 ENABLING TECHNOLOGY

The extent to which persons with impaired bodily functions are forced to live their lives differently than other people depends not only on therapeutic technology but also to a large part on a variety of other technologies, from wheelchairs to computer interfaces, from hearing aids to garage doors. Since the 1970s, handicap activists have urged us to see handicap less as a medical problem than as a consequence of social exclusion that is often mediated by technology. This standpoint was well expressed by Alison Davis:

> [I]f I lived in a society where being in a wheelchair was no more remarkable than wearing glasses and if the community was completely accepting and accessible, my disability would be an inconvenience and not much more than that. It is society which handicaps me, far more seriously and completely than the fact that I have spina bifida. (Quoted in [Newell, 1999, p. 172].)

It is important to observe the difference between a medical condition (such as being blind) and a social condition that it contributes to (such as being unable to read the newspaper). This can be expressed with the distinction between disability and handicap. Disability is an impairment of a bodily or mental function. Handicap is the presence of obstacles that persons with disabilities are subject to in society. Hence disability is inherent in the person, whereas handicap is a relation between a person and her environment [Amundson, 1992].

Technology with capacity to reduce the negative impact of having a disease or disability can be called *enabling technology* [Hansson, 2007c]. Leaving aside therapeutic technology that we have already treated, enabling technology can be divided into three categories: compensatory, assistive, and universal technology.

5.1 Compensatory and assistive technology

Compensatory technology is technology that replaces (fully or in part) a lost biological function by a new function of a general nature. Hence, whereas therapeutic technology reduces handicap by reducing disability, compensatory technology reduces handicap by providing new abilities that compensate for the disability. Some examples of compensatory technology are eyeglasses, hearing aids, speech synthesis systems, walking sticks, crutches, wheelchairs, orthotic appliances, ventilators, and equipment for total parental nutrition. Rehabilitation medicine that aims at replacing lost functions by new compensating ones makes much use of compensatory technology.

Assistive technology makes it possible for the individual to perform a task or activity despite an (uncompensated) disability or lack of function. Assistive technology provides abilities of a more specialized nature than what compensatory technology does. Typical examples are knives that require less strength than standard kitchen knives, plates and dishes that do not slide on the table, appliances for dressing, toileting, and bathing, remote controls for doors, windows, and light switches, textphones and videophones for the speech and hearing impaired, reading machines for the blind, etc. Adaptive interfaces of software products have become an increasingly important form of assistive technology, both for private life and on workplaces. However, the adaption of software has often lagged behind other technologies. As one example of this, many colleges and universities have ensured that handicapped persons have access to their buildings, but have failed to give them full access to their electronic information [Grodzinsky, 2000]. Household robots that assist disabled and elderly persons in a variety of daily activities are an important new development [Erlen, 2003].

Compensatory technology provides the person with general-purpose functions that can be used also in unforeseen situations, whereas assistive technology only provides solutions for more limited tasks. Therefore compensatory technology is more enabling than assistive technology. Hence, having a prosthesis that replaces a lost arm in a number of different tasks appears to be preferable to having a series of assistive appliances with which each of these tasks can be performed with only one arm.

5.2 Universal technology

Universal technology is technology that is intended for general use, not only for persons with a specific disease or disability. Without being restricted to persons with a disability, technology can be adjusted so that it includes them among its potential users. The difference between assistive and (adjusted) universal technology is often social rather than (in a restricted sense of the word) technological. Hence, a ramp that is used to enter a building both walking and in wheelchairs is universal technology; a wheelchair ramp at the back of the building intended only for those who cannot use the stairs at the front is assistive technology.

In the development of new technologies, accessibility for disabled persons is seldom treated more than at best as a side issue. Therefore, improvement or deterioration in terms of accessibility is often an unintended side effect of developments that have been driven by other aims. It is not easy to determine if the general trends in technological development are in general positive or negative for accessibility. There seem to be contradictory trends. One positive trend is mechanization that gradually decreases the need for physical strength in most occupations. Another positive trend is digitalization, that makes information more easily convertible to formats that are accessible to blind and deaf people [Cornes, 1993; Coombs 2003]. Mobile phones have also turned out to be more important for many handicapped people than for persons without a major handicap. A negative trend is increasing intellectual requirements, particularly on workplaces, that seem to be a consequence of many new technologies. This often makes life more difficult for mentally disabled persons. Hence, tentatively it seems as if ongoing technological developments make life easier for physically disabled persons but more difficult for those who are mentally disabled.

Appropriately adapted universal technology has the advantage over compensatory and assistive technology that it makes it possible for disadvantaged people to interact with the technological environment in the same way as others. As one example of this, if a machine — such as an elevator — has both visual and auditory signals, then both blind and deaf people can use it in the same way as people who see and hear. Similarly, if a heavy door is operated from a panel that is accessible from a wheelchair, then both walking and wheelchair-bound persons can open it in the same way. Therefore, universal technology is, as a general principle, superior to compensatory or assistive technology. It is therefore a plausible ethical standpoint that if a problem cannot be solved with therapeutic technology, then it should if possible be solved with universal technology, even if alternative solutions with compensatory or assistive technology are available.

However, contrary to therapeutic and compensatory technology, universal technology is usually not subject to decisions in the healthcare sector but rather in other sectors of society. This is in all probability a major reason why universal technology has often lagged behind therapeutic and compensatory technology.

6 GENERAL EFFECTS OF TECHNOLOGY IN MEDICINE

Technology has often been talked about very sweepingly in discussions on healthcare. In this chapter we instead focused on the impact on specific technologies and technological practices. However, there are some issues that do not relate to particular technologies but rather to the more general use of technology in healthcare. We will treat three major such issues: how technology shifts responsibilities, what effects it has on the quality of care and human contact, and whether it gives rise to unsound and perhaps unnatural dependence on artificial devices.

6.1 Shifting responsibilities

There are several ways in which current technological developments move responsibility for healthcare away from its traditional locus, i.e. physicians and nurses. The responsibility of companies that deliver medical equipment increases with the complexity of the equipment. In hospitals, bioengineers and clinical engineers take over some of the responsibilities of physicians, such as the calibration of advanced treatments. A quite different trend is the transfer of complex and sometimes life-critical equipment from the hospital to the patient's home, which confers more responsibility on patients and their relatives. Finally, as complex decisions are "delegated" to machines, some responsibilities become more diffuse, and bound to machines rather than to humans. Here, we will look more closely at the two last-mentioned of these trends, beginning with the shift of responsibility to patients and their relatives.

More and more patients receive treatments in their homes such as ventilator therapy and artificial nutrition through infusion pumps. These are treatments that were previously only administered in hospitals [Arras, 1994]. The increase in homecare is partly a response to patients' preferences, partly a response to economic pressures. "The combination of psychological benefits with cost containment makes home care seem an irresistible option" [Lantos and Kohrmann, 1992] (cf. [Kun, 2001]). Communications technology has an important role as facilitator of this development. Telemedicine allows for monitoring and diagnostics at home by the means of medical sensors connected to a personal computer. Temperature measurement, oximetry, electrocardiography, blood pressure measurement, and auscultation are among the diagnostic procedures that can be performed from a distance [Dansky *et al.*, 1999; Stanberry, 2000; Elger and Burr, 2000].

The administration of advanced diagnostic and therapeutic technology in homes has many advantages. When things go well in homecare, patients receive "the best of both worlds" [Arras, 1994], advanced medical treatment in the privacy of their own homes. Telemedicine in home care can be a way to ensure that access to healthcare is not limited by geographical location and ability to travel [Bauer, 2000; Elger and Burr, 2000].

However, technological homecare is not without its problems. For an increasing number of families, it has erased the boundaries between hospital and home, between intensive care unit and living room [Arras, 1994]. Sometimes parents and other relatives take over tasks that nurses perform only after taking special courses [Kirk, 2001]. Advanced technological home care with life-sustaining machines can place excessive burdens on family members, typically women, who live with a constant fear of failure. One of the most important ethical issues in home care is what tasks and responsibilities can and should be taken over by laypersons. "As home healthcare broadens to include traditionally hospital-based therapies, it is unclear whether traditional hospital norms, which place ultimate responsibility for decisions on professionals, or traditional home care norms, which place responsibility on parents, should apply" [Lantos and Kohrmann, 1992].

This can have negative social consequences. Homecare can make familiar domestic settings alien, and may confuse family roles. In comparison, hospitals can often allow patients greater autonomy, and therefore better preserve family relationships. Sometimes patients have a greater sense of privacy in hospitals than in homecare [Ruddick, 1994]. A patient's dependence on a spouse or a parent can be problematic for the relationship [Kohrmann, 1994]. Studies have shown stress and psychological problems in families who care for ventilator-dependent children at home [Lantos and Kohrmann, 1992; Arras and Neveloff Dubler, 1994; Kirk, 2001].

The other major shift in responsibilities emanates from a general tendency to automatize more and more advanced functions. Decisions are "taken over by machines" so that no human is directly responsible for them at the point in time at which they are made. Another way to express this is that decisions are predetermined in decision support systems.

Healthcare is often seen as one of the most promising areas for the introduction of computerized decision support. It has been shown in several cases that decision support systems can help the clinician in important ways, for instance by decreasing the risk of kidney failure, and providing more rapid treatment of critical laboratory abnormalities [Bates, 1997]. If a decision support system is connected to electronic patient records, it can include mechanisms for following up and for automated learning. Like other applications of artificial intelligence, an advanced clinical decision support system will therefore have capabilities in addition to those explicitly programmed into it.

We may very well be approaching systems in which computers perform what we usually see as the tasks of physicians: making diagnoses, performing therapies, and communicating with patients [Gell, 2002]. A system has already been tried out in which diabetes patients used a touch-tone telephone to obtain self-management instructions and dosage decision support from a computer. The result was encouraging; an improvement was shown in their diabetes management [Albisser, 2001]Nevertheless, important questions can be raised about the implications of such systems. If the advice was wrong, how important is it whether the patient communicated with a machine or with a human being? How can responsibilities be assigned when decisions are taken over by machines [Klieglis *et al.*, 1986; Hucklenbroich 1986]? Furthermore, what will the effects be on the physician–patient and nurse–patient relationships if much of the therapeutic-technical support comes from a machine whereas the psychological part of the support presumably stays with the physicians and nurses?

6.2 Technology, care and human contact

One of the most important effects of enabling technologies is to facilitate human communication. Hearing aids, textphones, appliances for reading and writing, speech reading programs, and various technologies for physical mobility are all examples of this. However, technology can also be used to replace human contacts or reduce the need for them. A phone call from a nurse can replace a personal

visit. A central electronic monitoring system can supersede assistant nurses at the bedside, and a nasogastric tube can be used instead of spoon-feeding.

In public debates, medical technology has often been accused of causing the dehumanization and depersonalization of healthcare and the objectification of patients. However, there is no inbuilt conflict between care and technology; technology can be used both in ways that improve care and in ways that make it less humane [Haber, 1986; Barnard and Margarete Sandelowski, 2001; Widdershoven, 2002]. In a balanced discussion on technology in healthcare it has to be realized both (1) that technology is not in itself dehumanizing and (2) that technology cannot replace genuine human contact and care.

For a practical example, we can consider the proposal to use virtual environments for training stroke patients. Virtual technology can be used to expose these patients to a wider range of sensory stimuli, over much longer periods, than what is otherwise possible in a hospital setting. This can yield benefits in terms of time and cost of therapy to stroke patients, who typically spend only 30-60 min per day in formal therapy. Thus, virtual reality "increases the possibility of stimulation and interaction with the world without increasing demands on staff time" [Wilson *et al.*, 1997]. However, potentially this technology can also be used to reduce individual, staff-to-patient contact. This is then a negative effect of the way in which the technology is used, not of the technology itself.

Recently, attempts to replace human contact with technology have in fact been made through the therapeutic use of companion robots. These products have been developed in Japan, where there is less resistance to robots with human features than in most other parts of the world. Elderly patients are invited to interact with robots such as the robot baby seal Paro that reacts when one speaks to it or pets its fur, and the "healing partner" Yumel that looks like a baby boy, has a vocabulary of 1200 phrases, and sings lullabies. Patients tend to appreciate these robots, cuddle with them and talk to them. Some patients with age-related dementia do not realize that they are interacting with a machine [Sullins, 2005]. Replacing human contact in this way is obviously problematic from an ethical point of view. It is questionable whether it is compatible with human dignity to provide demented patients with technological devices that they wrongly believe to be living beings. However, on the other hand, removing these robots without replacing them with true human contacts is not necessarily an improvement.

6.3 The technological imperative

Resistance to technological medicine has a long history. Around the year 1900 there was a "neohippocratic" movement among doctors who saw scientific medicine as a threat to the old art of medicine. One of the most prominent members of this movement was Ernst Schweninger, Bismarck's personal physician [Koch, 1985].

A much stronger such movement developed in the 1960s and 1970s as a counter-reaction to the rapidly growing use of mechanical and electronic equipment in healthcare. In 1968 economist Victor Fuchs introduced the term "technological

imperative", by which he meant the tendency to give the best care that is technically possible, even if its costs are high [Fuchs, 1968; Barger-Lux and Robert P. Heaney, 1986]. Much of the criticism of medical technology was couched in the term "medicalization". This term was used (and possibly invented) in 1961 by T. Szasz who originally used it to describe the incorporation into psychiatry of problems that should not be dealt with as psychiatric or otherwise medical [Nye, 2003]. The term was adapted by Ivan Illich (1926-2002), the foremost critic of technological medicine in this period. Illich, who has been incorrectly credited with inventing the term [Barnet, 2003, pp. 276 and 286], was an ardent critic of scientific medicine and in particular its use of technology [Illich, 1975]. In later years, the form of anti-medical movement that he represented has been significantly weakened.

Critics of medical technology have done great service to society by pointing out various problems in the use of this technology. However, much of their criticism is weakened by an (explicit or implicit) technological determinism: a belief that medical technology of necessity must have certain negative traits, such as being dehumanizing and standing in the way of good care. On the other hand, blind belief in the progress of medical technology can be equally dangerous.

BIBLIOGRAPHY

[Albisser *et al.*, 2001] A. Albisser, S C. Michael, S.C. En Chao, I.D. Parson, and M. Sperlich. Information technology and home glucose clamping. *Diabetes technology and therapeutics,* 3, 377-386, 2001.

[Alper and Beckwith, 1988] J. S. Alper and J. Beckwith. Distinguishing Genetic from Non-genetic Medical Tests: Some Implications for Antidiscrimination Legislation. *Science and Engineering Ethics*, 4, 141-150, 1988.

[Altenstetter, 2003] C. Altenstetter. EU and member state medical devices regulation. *International Journal of Technology Assessment in Health Care*, 19, 228-248, 2003.

[Altmann, 2001] J. Altmann. *Military Uses of Microsystem Technologies. Dangers and Preventive Arms Control.* Münster: agenda Verlag 2001.

[Amundson, 1992] R. Amundson. Disability, handicap, and the environment. *Journal of Social Philosophy*, 23, 105-119, 1992.

[Anderlik and Rothstein, 2001] M. R. Anderlik and M.A. Rothstein. Privacy and confidentiality of genetic information: What rules for the new science? *Annual Review of Genomics and Human Genetics*, 2, 301-433, 2001.

[Arras, 1994] J. D. Arras. The technological tether. *Hastings Center Report, Supplement* 24, S1-S3, 1994.

[Arras and Neveloff Dubler, 1994] J. D. Arras and N. Neveloff Dubler. Bringing the Hospital Home. Ethical and Social Implications of High-Tech Home Care. *Hastings Center Report, Supplement* 24, S19-S28, 1994.

[Bach and Barnett, 1994] J. R. Bach and V. Barnett. Ethical considerations in the management of individuals with severe neuromuscular disorders. *American Journal of Physical Medicine and Rehabilitation*, 73, 134-140, 1994.

[Balkany *et al.*, 1996] T. Balkany, A. V. Hodges, and K. W. Goodman. Ethics of cochlear implantation in young children. *Otolaryngology and Head and Neck Surgery*, 114, 748-755, 1996.

[Barash, 2000] C. I. Barash. Genetic Discrimination and Screening for Hemochromatosis: Then and Now. *Genetic Testing*, 4(2), 213-218, 2000.

[Barger-Lux and Heaney, 1986] M. J. Barger-Lux and R. P. Heaney. For better and worse: The technological imperative in health care. *Social Science and Med*icine, 22, 1313-1320, 1986.

[Barnard and Sandelowski, 2001] A. Barnard and M. Sandelowski. Technology and human nursing care: (ir)reconcilable or invented difference? *Journal of Advanced Nursing*, 34, 367-375, 2001.

[Barnet, 2003] R. J. Barnet. Ivan Illich and the Nemesis of Medicine. *Medicine, Health Care and Philosophy*, 6, 273-286, 2003.

[Bates, 1997] D. W. Bates. Commentary: Quality, Costs, Privacy and Electronic Medical Data. *Journal of Law, Medicine and Ethics*, 25, 111-112, 1997.

[Bauer, 2000] K. A. Bauer. The ethical and social dimensions of home-based telemedicine. *Critical Reviews in Biomedical Engineering* 28, 541-544, 2000.

[Behling, 2005] L. L. Behling. Replacing the Patient: The Fiction of Prosthetics in Medical Practice. *Journal of Medical Humanities*, 26, 53-66, 2005.

[Berreby, 1996] D. Berreby. Up with people: dwarves meet identity politics. *New Republic* 214(18), 14-19, 1996.

[Bertschinger *et al.*, 2008] D. R. Bertschinger, E. Beknazar, M. Simonutti, A. B. Safran, J. A. Sahel, S. G. Rosolen, S. Picaud, and J. Salzmann. A review of in vivo animal studies in retinal prosthesis research. *Graefe's archive for clinical and experimental ophthalmology*, 246(11): 1505-17, 2008.

[Booth, 1993] C. C. Booth. Clinical Research. In W. F. Bynum and R. Porter, eds., *Companion Encyclopedia of the History of Medicine*, pp. 205–299. Routledge 1993.

[Bowman, 2000] J. E. Bowman. Technical, Genetic, and Ethical Issues in Screening and Testing of African-Americans for Hemochromatosis. *Genetic Testing*, 4(2), 207-212, 2000.

[Brady, 1995] T. Brady. The Ethical Implications of the Human Genome Project for the Workplace. *International Journal of Applied Philosophy*, 10, 47-56, 1995.

[Bramstedt, 2003] K. A. Bramstedt. Contemplating total artificial heart inactivation in case of futility. *Death Studies* 27, 295-304, 2003.

[Bramstedt and Wanger, 2001] K. A. Bramstedt and N. S. Wanger. When withdrawal of life-sustaining care does more than allow death to take its course: The dilemma of left ventricular assist devices. *Journal of Heart and Lung Transplantation,* 20, 544-548, 2001.

[Büttner, 2002] J. Büttner. Naturwissenschaftliche Methoden im klinischen Laboratorium des 19. Jahrhundert und ihr Einfluß auf das klinische Denken. *Berichte zur Wissenschaftsgeschichte*, 25, 93-105, 2002.

[Coombs, 2003] N. Coombs. *Liberation Technology. Equal Access Via Computer Communication* 2003. http://ittimes.ucdavis.edu/spring2003/stories/past_ fall92_ liberation.htm.

[Cornes, 1993] P. Cornes. Impairment, Disability, Handicap and New Technology. In M. Oliver, ed. *Social work. Disabled People and Disabling Environments*, pp. 98-114. Jessica Kingsley Publishers 1993.

[Crouch, 1997] R. A. Crouch. Letting the deaf be Deaf. Reconsidering the use of cochlear implants in prelingually deaf children. *Hastings Center Report,* 27, 14-21, 1997.

[Dalsgaard Reventlow *et al.*, 2006] S. Dalsgaard Reventlow, L. Hvas, and K. Malterud. Making the invisible body visible. Bone scans, osteoporosis and women's bodily experiences. *Social Science and Medicine*, 62, 2720-2731, 2006.

[Daniels, 2000] N. Daniels. Normal Functioning and the Treatment-Enhancement Distinction. *Cambridge Quarterly of Healthcare Ethics,* 9, 309-322, 2000.

[Dansky *et al.*, 1999] K. H. Dansky, K.H. Bowles, and L. Palmer. How telehomecare affects patients. *Caring Magazine*, 18, 10-14, 1999.

[Davis, 1981] A. B. Davis. *Medicine and its Technology. An Introduction to the History of Medical Instrumentation*. Greenwood Press 1981.

[Deber and Leatt, 1986] R. B. Deber and P. Leatt. The multidisciplinary renal team: who makes the decisions?. *Health Matrix* 4, 3-9, 1986.

[Dolgin, 2001] J. L. Dolgin. Ideologies of Discrimination: Personhood and the 'Genetic Group'. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 32(4), 705-721, 2001.

[Elger and Burr, 2000] C. E. Elger and W. Burr. Advances in Telecommunications Concerning Epilepsy. *Epilepsia* 41, Suppl 5, S9-S12, 2000.

[Erlen, 2003] J. A. Erlen. Technology. Possibilities and Pitfall. *Orthopaedic Nursing*, 22, 310-313, 2003.

[Evidence-based Medicine Working Group, 1992] Evidence-Based Medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA*, 268, 2420-2425, 1992.

[Feinstein, 1967] A. R. Feinstein. *Clinical Judgement*. Krieger 1967.
[Feinstein, 1996] A. R. Feinstein. Two Centuries of Conflict-Collaboration Between Medicine and Mathematics. *Journal of Clinical Epidemiology*, 49, 1339-1343, 1996.
[Fielder, 1991] J. H. Fielder. Ethical issues in biomedical engineering: The Bjork-Shiley heart valve. *IEEE Engineering in Medicine and Biology*, 10, 76-78, 1991.
[Fuchs, 1968] V. R. Fuchs. The growing demand for medical care. *New England Journal of Medicine*, 279(4), 190-195, 1968.
[Gell, 2002] G. Gell. Safe, controllable technology? *International Journal of Medical Informatics,* 66, 69-73, 2002.
[Gin, 1997] B. R. Gin. Genetic discrimination: Huntington's disease and the Americans with Disabilities Act. *Columbia Law Review*, 97, 1406-1434, 1997.
[Goering, 2000] S. Goering. Gene Therapies and the Pursuit of a Better Human. *Cambridge Quarterly of Healthcare Ethics,* 9, 330-341, 2000.
[Green and Botkin, 2003] M. J. Green and R.J. Botkin. 'Genetic exceptionalism' in medicine: Clarifying the differences between genetic and nongenetic tests. *Annals of Internal Medicine* 138(7), 571-575, 2003.
[Grodzinsky, 2000] F. S. Grodzinsky. Equity of Access: Adaptive Technology. *Science and Engineering Ethics*, 6, 221-234, 2000.
[Haber, 1986] P. A. L. Haber. High Technology in Geriatric Care. *Clinics in Geriatric Medicine* 2, 491-500, 1986.
[Hansson, 2005] S. O. Hansson. Privacy, Discrimination, and Inequality in the Workplace. In S. O. Hansson and E. Palm, eds., *The Ethics of Workplace Privacy*, pp. 119135. Peter Lang 2005.
[Hansson, 2006] S. O. Hansson. Uncertainty and the Ethics of Clinical Trials. *Theoretical Medicine and Bioethics*, 27, 149-167, 2006.
[Hansson, 2007a] S. O. Hansson. What is Technological Science? *Studies in History and Philosophy of Science* 38, 523-527, 2007.
[Hansson, 2007b] S. O. Hansson. Praxis Relevance in Science. *Foundations of Science*, 12, 139-154, 2007.
[Hansson, 2007c] S. O. Hansson. The Ethics of Enabling Technology. *Cambridge Quarterly of Healthcare Ethics*, 16, 257-267, 2007c.
[Hansson, 2008] S. O. Hansson. Three Bioethical Debates in Sweden. *Cambridge Quarterly of Healthcare Ethics*, 17, 261-269, 2008.
[Haraway, 1991] D. Haraway. A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century. In D. Haraway, ed., *Simians, Cyborgs and Women: The Reinvention of Nature*, pp 148-181. Routledge 1991.
[Hofmann, 2003] B. Hofmann. Medicine as *Techne* – A Perspective from Antiquity. *Journal of Medicine and Philosophy*, 28, 403-425, 2003.
[Hucklenbroich, 1986] P. Hucklenbroich. Automatisation and responsibility. *Theoretical Medicine* 7, 239-242, 1986.
[Hudson, 2000] J. Hudson. What Kinds of People Should We Create? *Journal of Applied Philosophy,* 17, 131-143, 2000.
[Illich, 1975] Illich, *Ivan Medical nemesis: the expropriation of health*. Calder and Boyars 1975.
[Jacobson, 1998] N. Jacobson. The Socially Constructed Breast: Breast Implants and the Medical Construction of Need. *American Journal of Public Health,* 88, 1254-1261, 1998.
[Kirk, 2001] S. Kirk. Negotiating lay and professional roles in the care of children with complex health care needs. *Journal of Advanced Nursing*, 34, 593-602, 2001.
[Klieglis *et al.*, 1986] U. Klieglis, A.C. Renirie, and J. Schaefer. Medicus Technologicus. Reflections on the conflict between the physician's responsibility in decision-making and medico-technical automation. *Theoretical Medicine*, 7, 233-238, 1986.
[Koch and Lacquer, 1985] R. Koch and N. B. Laqueur. Schweninger's Seminar, *Journal of Contemporary History*, 20, 757-779, 1985.
[Kohrmann, 1994] A. F. Kohrmann. Chimeras and Odysseys. Toward Understanding the Technology-Dependent Child. *Hastings Center Report, Supplement* 24, S4-S6, 1994.
[Kröner, 2005] H.-P. Kröner. Äußere Form und Innere Krankheit: Zur klinischen Fotografie im späten 19. Jahrhundert. *Berichte zur Wissenschaftsgeschichte*, 28, 123-134, 2005.
[Kun, 2001] L. G. Kun. Telehealth and the global health network in the 21st century. From homecare to public health informatics. *Computer Methods and Programs in Biomedicine*, 64, 155-167, 2001.

[Lane and Bahan, 1998] H. Lane and B. Bahan. Ethics of cochlear implantation in young children: A review and reply from a Deaf-World perspective. *Otolaryngology and Head and Neck Surgery,* 119, 297-313, 1998.

[Lantos and Kohrmann, 1992] J. D. Lantos and Arthur F. Kohrmann. Ethical aspects of pediatric home care. *Pediatrics*, 89, 920-924, 1992.

[Launis, 2000] V. Launis. The Use of Genetic Test Information in Insurance: The Argument from Indistinguishability Reconsidered. *Science and Engineering Ethics*, 6, 299-310, 2000.

[Levy, 2002] N. Levy. Reconsidering cochlear implants: The lessons of Martha's Vineyard. *Bioethics,* 16, 134-153, 2002.

[Magnus *et al.*, 1996] J. H. Magnus, R. M. Joakimsen, G. K. Berntsen, A. Tollan and A. J. Søgaard. What do Norwegian women and men know about osteoporosis? *Osteoporosis International*, 6, 31-36, 1996.

[Maguire and McGee, 1999] G. Q. Maguire and E.M. McGee. Implantable brain chips? Time for debate. *Hastings Center Report,* 29, 7-13, 1999.

[Miles *et al.*, 1988] S. H. Miles, M. Siegler, D. L. Schiedermayer , J.D. Lantos, and J. La Puma. The total artificial heart. An ethics perspective on current clinical research and deployment. *Chest* 94, 409-413, 1988.

[Miller *et al.*, 2000] F. G. Miller, H. Brody, and K.C. Chung. Cosmetic Surgery and the Internal Morality of Medicine. *Cambridge Quarterly of Healthcare Ethics,* 9, 353-364, 2000.

[Mol, 2000] A. Mol. What Diagnostic Devices Do: The case of blood sugar measurement. *Theoretical Medicine and Bioethics*, 21, 9-22, 2000.

[Moore, 2000] A. D. Moore. Owning genetic information and gene enhancement techniques. *Bioethics*, 14, 97-119, 2000.

[Murray, 2002] T. H. Murray. Reflections on the ethics of genetic enhancement. *Genetics in Medicine* 4(6 Suppl), 27S-32S, 2002.

[Newell, 1999] C. Newell. The social nature of disability, disease and genetics: a response to Gillam, Persson, Holtug, Draper and Chadwick. *Journal of Medical Ethics*, 25, 172-175, 1999.

[Nunes, 2001] R. Nunes. Ethical dimension of paediatric cochlear implantation. *Theoretical Medicine,* 22, 337-349, 2001.

[Nye, 2003] R. A. Nye. The evolution of the concept of medicalization in the late twentieth century. *Journal of History of the Behavioral Sciences*, 39(2), 115-129, 2003.

[Pereira *et al.*, 2007] E. A. Pereira, A. L. Green, D. Nandi, and T. Z. Aziz. Deep brain stimulation: indications and evidence. *Expert Review of Medical Devices*, 4(5), 591-603, 2007.

[Persson and Hansson, 2003] A. P. Persson and S. O. Hansson. Privacy at Work – Ethical Criteria. *Journal of Business Ethics*, 42, 59-70, 2003.

[Ramesh, 1997] E. M. Ramesh. Time Enough? Consequences of Human Microchip Implantation. *Risk*, 8, 373, 1997

[Resnik, 2000] D. B. Resnik. The Moral Significance of the Therapy-Enhancement Distinction in Human Genetics. *Cambridge Quarterly of Healthcare Ethics,* 9, 365-377, 2000.

[Rhodes *et al.*, 1999] L. A. Rhodes, C. A. McPhillips-Tangum, C. Markham, and R. Klenk. The power of the visible: the meaning of diagnostic tests in chronic back pain. *Social Science and Medicine*, 48,1189-1203, 1999.

[Roche and Annas, 2001] P. A. Roche, P. A. and G. J. Annas. Protecting genetic privacy. *Nature Reviews Genetics*, 2(5), 392-396, 2001.

[Ruddick, 1994] W. Ruddick. Transforming Homes and Hospitals. *Hastings Center Report, Supplement* 24, S11-S14, 1994.

[Saha *et al.*, 1985] S. Saha, S. Misra, and P. S. Saha. Bioengineers, health-care technology and bioethics. *Journal of Medical Engineering and Technology*, 9, 55-60, 1985.

[Saha and Saha, 1997] S. Saha and P. S. Saha. Biomedical Ethics and the Biomedical Engineer: A Review. *Critical Reviews in Biomedical Engineering*, 25, 163-201, 1997.

[Silvers and Stein, 2002] A. Silvers and M. A. Stein. An equality paradigm for preventing genetic discrimination. *Vanderbilt Law Review* 55, 1341-1499, 2002.

[Stanberry, 2000] R. Stanberry. Telemedicine: barriers and opportunities in the 21st century. *Journal of Internal Medicine*, 247, 615-628, 2000.

[Sullins, 2005] J. P. Sullins. *Building the Perfect Companions: The Humane Design of Personal Robotics Technologies.* Manuscript, 2005.

[Veatch, 2003] R. M. Veatch. Inactivating a total artificial heart: special moral problems. *Death Studies*, 27, 305-315, 2003.

[Verweij and Kortmann, 1997] M. Verweij and F. Kortmann. Moral assessment of growth hormone therapy for children with idiopathic short stature. *Journal of Medical Ethics*, 23, 305-309, 1997.

[Vos and Willems, 2000] R. Vos and D. L. Willems. Technology in Medicine: Ontology, Epistemology, Ethics and Social Philosophy at the Crossroads. *Theoretical Medicine and Bioethics,* 21, 1-7, 2000.

[Warwick *et al.*, 2003] K. Warwick, M. Gasson, B. Hutt, I. Goodhew, P. Kyberd, B. Andrews, P. Teddy, and A. Shad. The Application of Implant Technology for Cybernetic System. *Archives of Neurology*, 60, 1369-1373, 2003.

[Warwick *et al.*, 2007] K. Warwick, M. N. Gasson, and A. J. Spiers. Therapeutic potential of computer to cerebral cortex implantable devices. *Acta Neurochirurgica*. Supplement, 97(2), 529-535, 2007.

[Wei *et al.*, 2001] F. Wei, G. D. Wang, G. A. Kerchner, S. J. Kim, H. M. Xu, Z. F. Chen, and M. Zhuo. Genetic enhancement of inflammatory pain by forebrain NR2B overexpression. *Nature Neuroscience,* 4(2), 164-169, 2001.

[White, 1999] R. J. White. Brain chip: Postpone the debate. *Hastings Center Report,* 29, 4, 1999.

[Whitehouse *et al.*, 1997] P. J. Whitehouse, E. Juengst, M. Mehlman , and T. H. Murray. Enhancing cognition in the mentally intact. *Hastings Center Report,* 27(3), 14-22, 1997.

[Widdershoven, 2002] G. Widdershoven. Technology and care from opposition to integration. In C. Gastmans, ed., *Between Technology and Humanity. The Impact of Technology on Health Care Ethics*, pp. 1262-1282. Leuven University Press 2002

[Wilkinson, 1993] L. Wilkinson. Epidemiology. In W. F. Bynum and R. Porter, *Companion Encyclopedia of the History of Medicine*, pp. 1262–1282. Routledge 1993.

[Willems, 2000] D. Willems. Managing one's body using self-management techniques: practicing autonomy. *Theoretical Medicine and Bioethics* 21, 23-38, 2000.

[Wilson, 1977] P. N. Wilson. Nigel Foreman, and Danaë Stanton. Virtual reality, disability and rehabilitation. *Disability and Rehabilitation*, 19, 213-220, 1997.

[Winkler, 1988] E. R. Winkler. Foregoing Treament: Killing vs. Letting Die and the Issue of Non-feeding. Pp. 155-171 in James En Thornton, *Ethics and Aging: The Right to Live, The Right to Die*, University of British Columbia Press 1988.

[Wolf, 1995] S. M. Wolf. Beyond 'Genetic Discrimination': Toward the Broader Harm of Geneticism. *Journal of Law, Medicine and Ethics*, 23, 345-353, 1995.

[Wolpe, 2000] P. R. Wolpe. Treatment, enhancement, and the ethics of neurotherapeutics. *Brain and Cognition,* 50, 387-395, 2000.

[Wood, 2002] J. Wood. The role, duties and responsibilities of technologists in the clinical laboratory. *Clinica Chimica Acta* 319, 127-132, 2002.

[Wood *et al.*, 2006] S. Wood, R. Jones, and A. Geldart. *The Social and Economic Challenges of Nanotechnology*. Swindon: Economic and Social Research Council 2003.

[Wooton, 2006] D. Wootton. *Bad Medicine*. Oxford University Press 2006.

Medicine, Health Care and Philosophy **4:** 335–346, 2001.
© 2001 *Kluwer Academic Publishers. Printed in the Netherlands.*

Scientific Contribution

On the value-ladenness of technology in medicine

Bjørn Hofmann
Center for Medical Ethics, University of Oslo, Box 1130, Blindern, 0318 Oslo, Norway (Phone: +47 22 84 46 45; Fax: +47 22 84 46 61; e-mail: b.m.hofmann@medetikk.uio.no

Abstract. The objective of this article is to analyse the value-ladenness of technology in the context of medicine. To address this issue several characteristics of technology are investigated: i) its interventive capacity, ii) its expansiveness and iii) its influence on the concept of disease, iv) its generalising character, v) its independence of the subjective experience of the patient. By this analysis I hope to unveil the double face of technology: Technology has a Janus-face in modern medicine, and the opposite of its factual face is evaluative.

Key words: ethics, technology, value-ladenness, values

Introduction

In order to address the issue of the value-ladenness of technology in the context of medicine, it is urgent to make clear what "value free" means.[1] "Value-free" apparently does not mean that something is free of being associated with values. There seems to be a general agreement that technology is related to issues of value. Technology has widely enhanced the possibilities of acting and producing which poses the question of how we *ought* to realise these possibilities (Schrader-Frechette & Westra 1997). Rephrased we might say that what *is* urges questions of *ought*. In this respect technology is part of the general question of what *the good life* is and clearly is associated to issues of value. Understanding value-ladenness as anything that poses value issues certainly answers the question of whether technology is "value-laden". It also replies to the question of how this influences medicine: by giving rise to a variety of ethical challenges technology makes medicine "value-laden".

However, this understanding of value-ladenness does not add to our theoretical knowledge of medicine.[2] Even proponents of "value-free" technology will agree that technology is associated with issues of value. In particular they argue that the values associated with technology are values of society at large (Bijker 1990; Hollander 1997; Tatum 1997), certain social classes (Rothman 1997) or particular interest groups (Vos 1991; Payer 1992; Moss 1991; Blume 1992).

Therefore in this study "value-free" will mean that values are aspects external to technology as such. Correspondingly, the claim that "technology is value-laden" will denote that values are related to technology qua technology. Technology does not only generate issues of value, but it is related to values as such. In other words, if technology is value-laden, it is not only a matter of what *is*, but also what *ought* to be, not only of what *could be done*, but what *ought* to be done.

Hence, the objective of this study is to analyse the value-ladenness of technology in the context of medicine. How then, can technology be conceived of as value-laden? There appears to be two major approaches to answer this question. The most common way to analyse the value-ladenness of technology is by an overall theoretical approach. There are several positions conceiving of technology as value-laden. It has been argued that technology represents an imperative enforcing humans to act in certain ways. Technology, under cover of being a mean, directs human ends and values. This position has been labelled *technological determinism* and its main issue is to investigate this *technological imperative* (Ellul, 1964; Winner, 1977; Smith and Marx, 1994).

From a phenomenological position it is claimed that technology is part of human understanding of being (Heidegger, 1953; Idhe, 1990). Man and his world are shaped by technology, which is of value not only as means for certain ends, but as a basic part of our being.[3]

An alternative approach to this theoretical analysis of value-ladenness of technology is to analyse technology's value-ladenness from a practical point of view: How do we recognise values of technology in medical practice? Instead of subscribing to any of the

mentioned monistic perspectives on technology and value, I will try to analyse how values are related to technology on a practical and detailed level. In other words, I will investigate whether the monistic theories of technology are adequate for analysing the issues of value. In particular, I will analyse a collection of well known examples to illustrate the wide range of value-ladenness related to medical technology. The examples will demonstrate how difficult it is to comprise technology's value-ladenness within a monistic theory. As a framework for this analysis I will investigate some key characteristics of technology in medicine. Technology is characterised as being:

i) *Interventive:* Through technology medicine has changed from assisting the healing capacity of nature to controlling and manipulating bodily healing itself.
ii) *Expansive:* Due to its interventive capacity technology has greatly expanded the field of medicine and increased its specialisation.
iii) *Defining disease:* By providing the basic phenomena to be studied and manipulated in medicine, technology strongly influences the concept of disease, and hence medical action. It defines what is diagnosed and what is treated.
iv) *Generalising:* It represents a general method for diagnosis, palliation and treatment. Its ability to generate reproducible results has made medicine a science.
v) *Liberating:* Technology has made medical knowledge independent from the subjective experience of the patient.

Hence, the objective is to investigate these characteristics in order to analyse the value-ladenness of technology in the context of medicine. In particular, it will be argued that technology does not only generate (external) issues of value, but it represents issues of values as such. Technology is value-laden on a constitutive level, which becomes particularly clear in medicine.

1. Interventive medicine

Hence, one of the main characteristics of technology in medicine is that it is interventive (*interveniere*). It has come to control and manipulate the organs, functions and processes of the human body. Conditions that earlier were fatal are today treated and cured. This interventive capacity of technology has greatly expanded the field of medicine, and it has changed medicine in several ways.

Firstly, whereas medicine earlier mainly was explanatory, it has now become manipulative. The function of humoral pathology was mainly to explain the observed phenomena. Practical measurement of and intervention with the processes of nature were of little interest (Hippocrates: *On ancient medicine*). The role of medicine was to explain and foresee the processes of nature. Today its function is to intervene in the observed processes. Practice comes before theory: Interventive methods are applied if they prove effective, independent of whether their mechanisms are known.

Secondly, the interventive capacity has altered the content of medicine. The explanatory entities of assistive medicine have been replaced by the manipulatory entities of technological medicine. Physiology, biochemistry and molecular biology have become basic subjects in medicine because they identify entities that can be manipulated. The interest, for example in the chemical substances of the human body, is due to the possibility of manipulating them. Hence, the interventive capacity of technological medicine has changed the subject matter of medical knowledge.

Thirdly, technological medicine has strongly influenced the classification of diseases. What is possible to manipulate and treat has been defined as a disease. The influence of technological medicine on the concept of disease will be dealt with later. Suffice it here to note that its interventiveness has influenced medical taxonomy. It influences what is and what is not subject to medical attention.

Fourthly, technology's interventive capacity has changed the status of medicine. Through the extended potential of action it represents power. The medical profession has gained power by the interventive and manipulative capacity of technology.

Altogether, the interventiveness of technology has altered medicine in a profound way, and this is an issue of value in several aspects.

Evaluative aspects of interventive medicine

This is not the place to enter into a discussion of the vast number of examples of evaluative challenges inherent in the *interventive capacity* of medicine. Only some issues will be investigated to illustrate the spectrum of fundamental evaluative issues: Firstly, technology challenges the concept of the patient. Secondly, it urges medicine to define its goals, and thirdly, to set limits to its activity. Additionally, there is an extended responsibility inherent in the extended potential of technological medicine.

The interventive capacity challenges the concept of the patient. It gives rise to the question: Who is the subject of the treatment – who is the patient? Technological medicine involves other subjects than

the traditional one-to-one patient-physician relationship. Transplant technology forces the physician to pay attention to the donor. Foetal surgery forces health care professionals to balance the concerns for the mother with the concerns for the child. In vitro fertilisation poses similar challenges. Perfusion of a brain-dead mother until her foetus is viable or of an anencephalic child until its vital organs can be transplanted into another baby represent similar types of evaluations. Xenotransplantation and cloning are other examples. These cases illustrate how technology challenges traditional values in medicine: the personal physician-patient relationship.

Moreover, the interventive capacity of technology challenges the goals of medicine (Kass, 1975; Hanson and Callahan, 1999). The case of life-sustaining treatment is a widely applied example. The possibilities for keeping comatose patients alive with respirators forced us to answer the question of *why*: What is the end of such treatment? Is it survival and extension of life, or is it the welfare of the patient? Inherent in issues of foetal surgery, human enhancement and genetic engineering there reside questions concerning the purpose of interventive treatment. The same questions are posed in cases where technological medicine is applied in excess, is futile, or is detrimental.[4] If the interventive capacities of technological medicine influence the actions and ends of medicine, they are issues of value. They do not only tell us what is, but also question what ought to be.[5]

Determination of the goals for interventive medicine touches upon an additional evaluative question: *Whose* goals? Does the interventive treatment serve the patient, the relatives, the professionals or society? The case of *hypoplastic left heart syndrome* might illustrate this (Bove and Lloyd, 1996; Hagemo et al., 1997; Kern et al., 1997). Here it is not obvious whether the complex, painful and risky treatment with low efficacy and effectiveness serves the benefits of the child, the parents, the skills of the professional or society. The difficulty of defining the goals of interventive medicine therefore relates to the concept of *who* is the subject in medical treatment. Hence, the interventiveness of technological medicine challenges patient autonomy.

Related to this urge for defining the goals of medicine due to technological interventiveness is a requirement to set limits to its activity. Where are the limits to what medicine should do? When the possibilities of treatment are substantially extended it becomes important to know when to abstain from or when to terminate treatment. Inherent in technology's interventiveness there is an issue of its limits, which is clearly displayed in medicine.

Additonally, the comprehensive capacity of interventive medicine is associated with an extended responsibility. The thalidomide case illustrates how the increase in interventive capacity of medicine also increases the seriousness of its consequences if applied erroneously. An increase in the possibility of doing good also enhances the potential of doing wrong. The extensive possibilities related to technological medicine lead to extended responsibilities.[6]

So, as a result of the interventive capacity of technological medicine, the concept of *patient* in medicine is challenged. Due to the increased interventive capacity the goals and limits of medicine have to be redefined, and physicians face an enhanced responsibility. Altogether, what is possible in technological medicine is related to the questions of what *ought* to be done. *Can* implies the question of *ought*. Hence, inherent in the interventive capacity of technology in medicine we encounter issues of value. Inherent in factual issues of *how* to do things, there is an evaluative question of *if* and *what* to do. The new possibilities force us to cope.

2. The technological expansion of medical knowledge

Related to the expanded possibility to intervene, there is an expanded possibility to know. Due to the interventive capacity and the widespread application of technology, the *Corpus Medicorum* has become more extensive and specialised than ever.

This has given rise to a set of demanding questions: Is the new knowledge *good or bad*? Furthermore, *how* is this comprehensive knowledge to be applied? For example, is it right to clone humans, or to make hybrid pigs for xenotransplantation? How shall we ration technological medicine? It has been argued that the evaluative aspects of this expansion of medical knowledge have been ignored (Jonas, 1985; Gadamer, 1993) and, as a consequence, that medicine does more harm than good (Illich, 1975; Lewis, 1977; Stewart-Brown and Farmer, 1997; Sharpe and Faden, 1998; Fischer and Welch 1999). Is it true that we have grown to become technological giants, while we are still to be considered as ethical embryos? Science and technology does not appear to liberate medicine from ethical issues, on the contrary: "It is paradoxical, perhaps, that to apply the creations of our newest scientific disciplines, physicians must reexamine the moral principles by which they act, and turn to ethics, one of our oldest humanistic disciplines" (Reiser, 1977, p. 55).

It is beyond the scope of this study even to sketch the features of this technologically determined expansion of medical knowledge. Only the case of predictive testing will be employed to exemplify the expansion of medical knowledge and its evaluative challenges.

Predictive testing – a case study

Particular to predictive testing is that it can be used to detect cases of disease where the patient has no subjective experience of being ill. Such *asymptomatic diseases*[7] seem to be rich in evaluative consequences. The aims of treatment are altered from removing causes and symptoms of experienced illness to treating unperceived disease. This represents a fundamental epistemological and evaluative change in medicine. Epistemologically, medical knowledge seems to be independent of the patient's subjective experience. This will be discussed in detail later. Evaluatively, the initiative of care and cure is shifted from the patient seeking help to the health care provider offering assistance.[8] Hence, medicine seems to have liberated itself from the initial initiative of the patient.

It has transgressed its traditional ethical basis where a person seeks help because of pain, discomfort, weakness, or ailment. Furthermore, medicine's independence of the patient's illness gives health care unrestricted power to prescribe treatment. Misuse of such power is not difficult to imagine, and how to manage this power is obviously an evaluative challenge. Predictive diagnostics, therefore, represent a change in the ethical status of the patient.

Additionally, some cases of *asymptomatic diseases* would never have become apparent to the patient if they had not been detected by a predictive test. The patient would never have developed symptoms during his or her lifetime. (Black and Welch, 1993; Stewart-Brown and Farmer, 1997; Kevnanagh and Broom, 1998). Papillary carcinoma of the thyroid, ductal carcinoma in situ of a woman's breast and adenocarcinoma of the prostate are examples of such cases.[9] So far, there is no way of predicting who will develop symptoms and who will not. If all the detected instances were followed up therapeutically, more healthy persons would be treated. Predictive testing, hence, increases the prevalence of the disease. Whether it is *good* or *bad* for medicine to "make people diseased" in this manner is a question of value.

Correspondingly, knowledge of a detected disease may make a person anxious and ill. The uncertainty related to this kind of medical knowledge may have a negative physical and psychosocial effect.[10] It has been shown that technological markers, e.g. foetal ultrasound, can result in anxiety and can have a negative influence on health (Malone, 1996). In this respect the technological expansion of medical knowledge can be harmful.[11] This illustrates the evaluate aspects related to new knowledge, which is especially important with diagnostic methods where no treatment exists for the detected disease.

Furthermore, predictive tests embody the evaluative issue of how much pain and inconvenience a person should be exposed to in cases where the probability for a disease developing is small. Is it right to remove the colon of a patient who has a hereditary polyposis and a mutation of the APC-gene (Ponder, 1997)? There is a profound difference between a person who is ill and needs help and a person who is not ill, when it comes to exposing them to treatment and the related pain and risk (Skrabanek, 1994, p. 36).

Altogether, predictive tests can make people diseased. Firstly, they can define people who do not feel ill as diseased. Thus they transgress the initiative of the patient. Secondly, they might lead to treatment of persons who never in their lifetime would have developed symptoms. Thirdly, the knowledge of an unperceived disease may make people both ill and diseased. They force us to deal with risk and uncertainty. Hence, predictive tests represent a *medicalisation* of human conditions. At what level we will allow this to happen is not a purely factual matter, but a matter of values as well.

Epistemic insufficiency

One of the difficulties due to this technological expansion of medical knowledge is, as argued, knowledge of disease without illness. But the opposite situation might also be problematic: where the patient is ill, but no disease can be detected. Is the patient then not diseased? Does he not qualify for treatment or care? If he does, by what means? Is he socially, but not medically diseased (Räikkä, 1996)?

Cases of illness without disease equally represent basic evaluative challenges to technological medicine. Despite the impressive amount of medical knowledge in ever more specialised sub-domains they illustrate an *epistemic insufficiency* in medicine. The knowledge of technological medicine is imperfect (Thomas, 1977). "There is a vast ocean of ignorance at the heart of medicine" (Le Fanu, 1999 p. 178).[12] This does not, however, differ from other systems of medical knowledge. All theoretical frameworks of medicine seem to be insufficient. The difference is that technological medicine appears to be *omnipotent* and *omniscient*. If the limits of medical knowledge are not acknowledged, many patients may suffer. Thus, ignorance of the *epistemic insufficiency* appears to be an issue of value. Ignoring the *docta ignorantia* in technological medicine is a matter of *good* and *bad*.

In addition there is a high turnover of medical knowledge. Yesterday's method is out-dated today. This turnover pushes the evaluative questions forward: What knowledge is *good* and how ought it to be applied? Is it immoral not to offer patients help

according to the most up-to-date knowledge? In particular it raises a practical question highly relevant for clinicians: How is it possible to be updated? When is the right time to change to a new method? How much better must a new method be before its benefits outweigh the costs of abandoning a well-established method? How are we to evaluate the efficacy, effectiveness and efficiency of new methodology?

Furthermore, technological medicine presents more possibilities for diagnosis and treatment than available resources can realise. Thus technological medicine has enhanced the problem of triage and forced us to ration recourses (Reiser, 1978; Aron and Schwartz, 1984; Anspach, 1987; Rothman, 1997). Some of the patients with diseases that can be detected and treated will not receive treatment. Which patients are to be given a heart-transplant? Who shall be treated for cataracts or have dialysis and who shall not have? The questions of *whom* shall be given health care services and *who* is to decide are practical and evaluative questions. They cannot be answered by simply referring to the descriptive powers of technology or resolved by implementing more technology.

Hence, the technological expansion of medical knowledge includes evaluative challenges. Knowledge of *how* the human body works and reacts, and *what* to do to influence it, comprises the question of *when* and *how* this knowledge *ought* to be applied and when to recognise its limits.

3. The technological constitution of disease

Technology appears to have become a paradigm in medicine by prescribing ways of detecting, identifying and treating disease. Disease now can be measured with objective instruments (Twaddle, 1993, p. 9). *Epilepsy*, originally conceived as a spiritual influence (Hippocrates: *The sacred disease*), through technology (electroencephalography, microscopic techniques, chemical analysers) has become a disturbance of electrical activity of the brain caused by paroxysmal malfunction of cerebral nerve cells. In the same manner a variety of cardiac conditions are defined by specific ECG-patterns, ultrasound flow measurements and radiographical morphology. The ability to measure blood pressure and to identify *Helicobacter pylori* has made such signs and markers define disease.

The technological influence on the concept of disease is not, however, limited to diagnosis. The success of technology in medicine has made technology the criterion of demarcation for treatment (Brown, 1985, p. 317). The methods of technology determine what is treatable and thereby set a precedent for what is to be treated.[13] Medical technology has become the measure of all things; a kind of *ars mensura*, or a *technê metriké*[14] of the modern age, being the measure of what is good and bad, what is diseased and what is not diseased, what is to be treated and what is not to be treated.

Therapeutically, the technologies of corrective surgery, blood pressure regulation and artificial fertilisation have made health care professionals treat these conditions as diseases: *hypoplastic left heart syndrome, hypertension* and *infertility*. Decisions and prognosis have come to be based on technology (Anspach, 1987; Tijmstra, 1989). Mitcham elegantly summarises this influence of technology on concepts of medicine:

> Medicine is increasingly defined ... by the type and character of its instruments (from stethoscope to high-tech imaging devices) and the construction of special human-artefact interactions (synthetic drugs, prosthetic devices). Indeed, the physician-patient relationship, medical knowledge, and the concept of health are all affected by technological change. (Mitcham, 1995, p. 2477).

Technology is not only involved in defining disease, but also in generating knowledge of disease. It has become the *definiens* of disease and appears to have become the paradigm method of medicine. Technology constitutes the categories of the medical gaze. "The technology mediates between the seer and the seen and what is seen becomes largely constituted by technology. This is why practices change with the development of new technologies" (Cooper, 1996, p. 394). Advances in technology facilitate the identification of new markers that will be treated as disease (Whittle, 1997). Technology comprises the physiological, biochemical and bio-molecular objects and events that constitute the disease entities in both diagnostics and treatment. For example, angiography, echo-doppler and tissue-velocity-imaging have resulted in an extended classification of myocardial infarction. Thus, epistemologically, ontologically and practically, technology is involved in constituting the concept of disease.

Technology, disease and value

Does this technological constitution of disease mean that technology has enabled a descriptive conception of disease? This does not seem to be the case. As previously argued, the interventive capacity of technology and its expansion of medical knowledge is not able to transcend issues of value. The concept of disease will be subject to the same evaluative challenges as the technology that defines it. Some of these

have already been discussed. However, other evaluative aspects appear to be related to the technological constitution of disease as well.

Defining disease by setting limits to what is normal and what is pathological is a matter of value (Canguilhem, 1991). Although technology offers a method of reproducible detection and identification of diabetes, defining the limits of normality is nevertheless an evaluative issue. The limits of diabetes defined by the American Diabetes Association (ADA) or by WHO are not factual descriptions. If one applies the WHO limit instead of that of ADA, then the prevalence of the disease is almost doubled (Wahl et al., 1998). Hence, the WHO definition of diabetes makes people diseased. The definition of normality, and thus disease, is an evaluative matter (Robinson and Bevan, 1993).

Furthermore, the sensitivity to the markers used to detect disease is continuously improved, as technology develops. This *increased sensitivity* expands the range of conditions qualifying for the status of disease. Thus, technology lowers the limits of disease and increases its prevalence. The detection of increasingly milder cases results in treatment of an increasing number of conditions. In practice technologically increased sensitivity results in a *lowered treatment threshold*. Increased sensitivity and lowered treatment comprise the evaluative issues of what is *good* diagnosis and what is *good* treatment. They include issues such as futile treatment and medicalisation (Fischer and Welch 1999).[15]

Moreover, technology has altered the end-points of medical activity. Technology defines the entities and markers to be studied and manipulated. In practice it tends to make medicine pursue *soft end-points* like *cardiac blood flow* and *cholesterol concentration*, and constitutes such conditions as diseases. When these markers are within normal limits, the patient is per se healthy.

However, the selection of end-points is a matter of value, and manipulating soft end-points does not guarantee results in terms of *hard end-points* such as survival and morbidity. Clinically the prevalence of prostate cancer in men aged between 60 and 70 is about 1%. However, by applying transrectal ultrasound or MRI more than 40% of men in the same age group have been diagnosed as having prostate cancer (Monti et al., 1989). Technology's focus of attention is on diagnostic and therapeutic impact and not on patient outcome (Bruke, 1994; Pickering, 1996). This technological affinity to soft end-points can be conceived of as a form of medicalisation and a form of disregard of patient autonomy.

Thus, inherent in the technological constitution of disease the measure of disease is changed, the limits to normality must be set and the prevalence of disease and the outcome of treatment are altered. Hence, the *technological constitution of disease* is a matter of value. It influences who is diseased and who is not, who is entitled to treatment and who is not, who will receive economic support, and who will not.

The objective here was neither to give a detailed description of a technological conception of disease, nor was it to give an exhaustive analysis of the evaluative issues of the disease concept. More modestly, the objective was to argue that the conception of disease is influenced by technology and that this reveals its value-ladenness. The issues of value cannot be removed from a technologically constituted concept of disease.

4. Generalising technology

One important characteristic of technology is its generalising ability. Technology facilitated the study and identification of the general in the particular. The ECG and X-ray rendered an objective way to scrutinise disease.

> Ophthalmoscope, broncoscope, etc. allow him [the physician] a direct view of the conditions of many parts. Experimental medicine enables the physician to interpret his findings so as to translate the language of symptoms and tests into the language of physiological processes. Here then is a scientific approach to individual sickness (Temkin, 1963, p. 636).

Technology eliminated both the singularity of the patient and subjectivity of the physician (Reiser 1978) and strongly influenced the postulates of causation in medicine (Evans, 1991). In short, technology made medicine a science (Temkin, 1963; Cassell, 1993, p. 38).

Technology facilitates the translation of individual illness into the objective language of physiology (Ferkiss, 1969; Jonsen, 1990, p. 25).[16] Through technology medicine gains objective data (Jonsen, 1990, p. 25), and technology represents a standard method of detection, identification and treatment of disease. In this way technology accounts for the reproducibility of results and for the accumulation of nomological knowledge. The MRI-machine presents a standard image of the human brain and automated laboratory analysers produce positive test results when the number and shape of blood cells deviate from normal statistical values.

This abstracting and generalising characteristic has been crucial for the argument that technological medicine is value-neutral (Sundström, 1998). Nevertheless,

rather than escaping the evaluative, the generalising attribute of medicine emphasises its value-ladenness. This value-ladenness can be illustrated by scrutiny of some of the flaws of this generalising characteristic.

Evaluative aspects of generalising technology

Let me briefly mention four flaws due to technological generalisation frequently referred to in the literature and then investigate some of the value related issues. Firstly, technological generalisation is based on populations rather than on the individual. The single patient might gain from general methodology, but might also suffer from it, due to natural variation in a population (Jonas, 1985; Gadamer, 1993, Delkeskamp-Hayes and Cutter, 1993).

Secondly, no technological method is absolutely effective, nor perfectly accurate and reliable. The same blood sample tested with the same chemical analyser may give different results for consecutive tests, e.g. blood gas measurements. There is statistical variation in the results due to the technological method. This might lead to erroneous diagnosis and treatment. The test can fail to detect disease and can detect disease when there is none.

Thirdly, inter-observer and intra-observer variability reduces the effectiveness of the method. Even if there was no variation in the population and the method was perfectly accurate and reliable, there would still be variation in the application of diagnostic and therapeutic technology. Different physicians apply technology differently in different cases (Jennett, 1988; 1994). Hence, the practical implementation and particular application of even a perfect method might be flawed.

Fourthly, technology is applied to different populations than the one they are tested on. Obviously tested technology is not applied to the test population again. This calls for careful judgement. It is well-known that diagnostic procedures and types of treatment that have been tested on hospitalised patients have been applied in general practice, and methods tested on men have been applied to women, which has resulted in erroneous diagnosis and treatment.

These profound flaws of the technology of medicine present evaluative challenges. On a general basis it is argued that the generalised method in medicine is erroneous (Gorovitz and MacIntyre, 1976, Leape, 1994). How we handle this inherent error in medicine is a matter of *value* and not only of *fact*. Let me briefly investigate some of the evaluative aspects.

Firstly, the question of how we handle the insufficiency of the generalising technology is an evaluative matter. How many false positives and false negatives will we allow? What level of significance do we accept? How much are we willing to let some patients suffer to help others? What responsibilities do health care professionals have towards the healthy persons that are treated and the diseased persons who are ignored? The very definition of confidence intervals is evaluative and the concepts of false negatives and false positives are issues related to *good* and *bad*.

Secondly, the ability to communicate the possibilities and restrictions of medicine due to its generalisation relate to ethical matters such as patient autonomy, informed consent and paternalism. Does the patient understand the uncertainty and risk? How do we act if he does not?

Thirdly, it has been claimed that the generalising method of technology in medicine tends to alter the physician's responsibility for the individual patient (Jonas, 1985; Gadamer, 1993, Delkeskamp-Hayes and Cutter, 1993). It is accused of freeing the physician from personal obligation towards the patient. "Western medicine and the modern paradigm of knowledge are heavily biased towards abstraction, we all tend to feel drawn away from the attempt to identify with the patient's experience" (McWhinney, 1997).

In other words, generalisation by technology leads to what might be called an *epistemic abstraction* from the particular patient, which has adherent evaluative aspects. Whether this *epistemic abstraction* also results in a corresponding *evaluative abstraction* from the patient will be discussed in the following section. The point here is that the generalising characteristic of technology does not make medicine escape issues of value. Handling the *epistemic abstraction* and its flaws is not a matter of how nature *is*, but of how we *ought* to live. The technological generalisation in medicine is in itself an evaluative matter.

5. Technological emancipation from the subjective patient

A crucial aspect of the technological generalisation discussed above is its abstraction from the individual person. Technology has altered the relationship of medicine to its subject matter: the patient. In other words, the objectivity of medicine is achieved by making the patient an object and liberating itself from the patient's subjective experience. However, this independence from the patient is an evaluative issue.

It is argued that before the Eighteenth Century, medicine was based on the patient's narrative of his or her symptoms. In addition to this subjective portrait of the illness, the physician observed the patient's appearance and behaviour as well as any signs of disease. During the Eighteenth and Nine-

teenth Centuries medical instrumentation enabled and extended the physical examination of patients, which made the physician less dependant on subjective narration (Reiser, 1995, pp. 1–90). The stethoscope gave the physician direct access to the disease. Measuring blood pressure gave an objective measure of internal conditions in the patient. The introduction of machines such as the ECG, X-ray and chemical laboratory analysers during the Nineteenth and Twentieth Centuries further enhanced the objectivity of medicine (Reiser, 1995, pp. 91–157). In addition to removing the subjective errors introduced by the patients, technology also reduced the number of erroneous judgements made by physicians. Technology liberated medicine from the subjective, individual and emotional factors, which confused the conception of the real objective disease. "Twentieth-century technology with all its progress had tended to push the human dilemmas of illness out of the doctor's thoughts, and replace them with laboratory facts derived from tests on the patient's body" (Reiser, 1978, p. 225).

Due to the generalisation in medicine the individual patient today contributes to the *Corpus Medicorum* only as one of many. The epistemic significance of the individual is reduced to a statistical entity. Accordingly, technology creates a physical distance between the physician and the patient (Jennett, 1994, p. 862), making it a 'stranger medicine' (Veatch, 1085; Rothman, 1991).

"Technological methods move the evidence employed in diagnosis away from the patient and reduce the impact of the patient's particularity on the physician" (Cassell, 1993, p. 36). The capacities of technological medicine have excluded the individual patient as the epistemic basis of medicine (Le Fanu, 1999 p.194). The essential question following from this is whether the evaluative status of the patient has been altered correspondingly.

Critics of modern medicine claim that technology's focus on the objective and the general has resulted in a neglect of the individual patient (Glover, 1977; Pellegrino, 1979; Jonas1985; Cassell, 1993; Gadamer, 1993). This transgresses the traditional normative basis of medicine. Ever since the awakening of medical self-consciousness, the raison d'etre of medicine has been to heal and help the individual patient.[17] The objective of medicine was the *good* of the particular patient. With technology in medicine there has been "a detachment from the suffering of [the] patient" (Cassell, 1993, p. 34). This is a detachment of the professional from the personal, disease from illness and signs from symptoms, making medicine face profound evaluative challenges such as medicalisation, reductionism, curative bias and paternalism. As already mentioned, there is a shift in initiative due to technology: the patient does not seek the health care system because he or she feels *bad*, but because the technological method detects something that is considered to be *bad* for the patient. The evaluative initiative is shifted from the patient to the health care system.

Hence, there appears to be a reduction of the evaluative status of the patient corresponding to the reduction in epistemic significance; there is an *evaluative abstraction* from the patient matching the *epistemic abstraction*. This represents what might be called an *evaluative ignorance of the individual* in technological medicine.

Evaluative characteristic of technological medicine

Altogether, the technology of medicine has been characterised by the following attributes:

i) *Interventive capacity:* Taking on an interventive and manipulative attitude.
ii) *Epistemic expansion:* The substantial extension of *Corpus Medicorum* due to technology.
iii) *Constituting disease:* The influence of technology on the concept of disease.
iv) *Generalising:* The technological generalisation of medical knowledge.
v) *Liberating from the subjective experience of the patient:* Making medical knowledge independent of the subjective experience of the patient.

The practically oriented analysis of these characteristics has revealed their inherent evaluative aspects. Within the possibilities of technology resides the question of whether it is *good* or *bad* to realise them. In concert with the potential of technology we face issues of *how, when, why, for whom*, and *by whom* it is to be applied. Within the knowledge of what *is* and what *can* be done with medical technology resides the challenge of what we *ought* to do. At the same time as technology expands our potential for action it urges us to define the ends of and set limits to its application. The relationship between technology and value comes particularly clear in medicine, explicitly dealing with issues of *good* and *bad* of the body (and mind).

In this study I have not dealt with the details on how in particular values relate to technology. This is the issue of another study. Here the main objective has been to argue that there is a close relationship between technology and value, particularly apparent in medicine. In other words: there is a close relationship between technology and ethics. Technology represents a Janus-face in medicine. The opposite of technology's descriptive face is evaluative.[18]

Concluding remarks: The Janus-face of medicine

The investigation of the relation between technology and value seems to be rich in consequences. Firstly, it is apparent that technology does not exclusively represent value-neutral means towards an external end. The study seriously questions the commonplace value-neutrality dictum.[19] The evaluative challenges related to technological medicine are not issues of conflicting external ends and cannot be resolved by agreeing upon external goals of medical activity. Technology, being inherently evaluative, constitutes medical knowledge. Technology makes medicine a scientific, but also a moral enterprise.

Secondly, even though the study has made me question the value-neutral dictum of technological medicine this is done without subscribing to one of the monistic theories of technology. The examples illustrate a wide range of value-ladenness of technology in medicine and demonstrate the difficulties of subscribing them all to one of the traditional critiques in the philosophy of technology. The monistic theories appear to fail to comprise the vast variety of value-aspects of technology in medicine. Additionally, the analysis shows the fruitfulness of a detailed approach to medical practice.

Thirdly, medicine is particularly suitable to study the value-ladenness of technology because its evaluative aspects are easily recognisable. Issues of value are widely recognised in medicine, and (bio)medical ethics is an important branch of moral philosophy (Toulmin 1986).

Hence, the conclusion of the study can be phrased: "*is* implies *ought*", but in the sense that the matter of what *is* in medicine comprises the evaluative issue of how it *ought* to be. There is reciprocity between *is* and *ought*; between the possible and the actual; between knowledge and its application; between fact and value. That is, there is a constitutive relationship between values and technology in medicine. By stepping into the doorway (*januae*) of technology we are already in the realm of value.

Notes

1. There appear to be many kinds of value: economic, esthetical and moral. To restrict the topic, "value" will in this study refer to moral value.
2. Value is not related to technology as such, but in the same manner as value relates to other objects and actions: they can be of value.
3. In the philosophy of medicine we can recognise both the position of technological determinism (Bennett, 1977; Hellerstein, 1983; Tijmstra, 1989; Cassell, 1993; Davidson, 1995; Muraskas et al, 1999) and the phenomenological approach (Cooper, 1996).
4. In particular, see (Illich, 1975; Reiser, 1978; Jennett, 1986; Payer, 1992, pp. 37–52; Cassell, 1993; Schneidermann et al., 1995; Tijmstra, 1989; Fischer and Welch, 1999).
5. Screening is a case that further exemplifies the difficulties of defining goals of medical treatment (Black, 1993; Stewart-Brown and Farmer, 1997; Kevnanagh and Broom, 1998; Kerbel et al., 1997; Whittle, 1997; Malone, 1996; Chevenak, 1998). The benefits of discovering disease have to be weighed against their costs, such as medicalisation of people, false positive or false negative results, detection of cases that are untreatable, anxiety among patients, and application of technological methods by doctors who lack clinical competence. The task of weighing the ends involved in such complex situations is certainly an evaluative matter.
6. The substantial increase in malpractice suits may be an indication of this.
7. Cases of detected disease without any symptoms have also been called lanthanic diseases (Feinstein, 1967).
8. Cases of health care where patients do not request help have been called non-iatropic diseases (Feinstein, 1967). Such cases seem to be of ethical relevance in profit maximising health care systems appealing to people's uncertainty, anxiety and concern for their health.
9. Cases of detected disease that would never have become appearent to the person have been called pseudodiseases (Helman, 1985; Fisher and Welch, 1999, p. 449).
10. See for example (Tijmstra, 1989; Green, 1990; Black and Welch, 1993; Kevnanagh and Broom, 1998).
11. The way that technological knowledge may be harmful can be called technological stigmatisation.
12. The incompleteness of medical knowledge is also demonstrated by the fact that a large number of diseases have unknown aetiology. In many cases medicine can only treat the symptoms and not the causes.
    However, can technological medicine ever reach complete knowledge? Gorovitz and MacIntyre argue that medical knowledge will always be incomplete, and that ignorance of this fact makes medicine errenous (Gorovitz and MacIntyre, 1976). Gadamer also argues that there is an epistemic insufficiency in technological medicine. "Aber trotz allen Fortschritten, die die Naturwissenschaften für unser Wissen um Krankheit und Gesundheit gebracht haben, und trotz dem enormen Aufwand an rationalisierter Technik des Erkennens und Handelns, der sich auf diesem Gebiete entfaltet hat, ist der Bereich des Unrationalisierten hier besonder hoch" (Gadamer, 1987, p. 259). Correspondingly, Paul argues that there is a theoretical insufficiency due to a gap between theory and practice in medicine, termed "Hiatus theoreticus". This is an epistemological void typically inherent in the stock of medical knowledge itself (Paul, 1998, p. 247).
13. The technological focus on treatment has contributed to what has been called the curative bias in modern medicine, which also is rich in normative consequences.
14. See (Gorgias 356d4–e2).
15. Among these are cases that would otherwise have healed by themselves (*trivia*).

16. For example, the stethoscope enabled the physician to listen to sounds from vessels. The classification of these sounds (Korotkoff) gave a general method of measuring blood pressure. This facilitated the correlation of blood pressure and certain pathological states.
17. See (Hippocrates: *The oath; On the art* III). Both Plato and Aristotle recognised that the challenge in medicine was not the content of medical knowledge, but how it should be applied in particular cases (*Phaedrus* 268a7–c4; *Nicomachean Ethics* 1104a4–6; 1137a10–25; 1097a11–4; 1143b18–32; 1180b5–23).
18. Temkin discusses the "Janus-face" of medicine in the context of the history of medicine (Temkin, 1977). The one face looks into the past, enabling the other to view into the future of the profession. In this study the concept of 'the Janus-face of medicine' is applied to emphasise the relationship between medical technology and ethics. The one face looks into the world of how things *are*, the other how they *ought* to be.
19. In the philosophy of technology the value-neutrality dictum has also been characterised as the voluntarist position (Winner, 1977, pp. 53–54; 60–63; 76–77).

References

Anspach, R.R.: 1987, 'Prognostic Conflict in Life-and Death Decisions: The Organization as an Ecology of Knowledge', *Journal of Health and Social Behavior* 28, 215–231.

Aron, H. and Schwartz, W.: 1984, *The Painful Prescription: Rationing Health Care*. Washington DC: Brookings Institution.

Bennett, I.L.: 1977, 'Technology as a Shaping Force', in: J.H. Knowles (ed.), *Doing Better and Feeling Worse*. New York: Norton & Co, pp. 125–133.

Bijker, W.E.: 1990, *The Social Construction of Technology*. Cambridge MA: Ph.D. Thesis, MIT.

Black, W.C. and Welch, H.G.: 1993, 'Advances in Diagnostic Imaging and Overestimation of Disease Prevalence and the Benefits of Therapy', *New Engl. J. Med* 328, 1237.

Blume, S.S.: 1992, *Insight and Industry. On the Dynamics of Technological Change in Medicine*. Cambridge Mass: The MIT Press.

Bove, E.L. and Lloyd, T.R.: 1996, 'Staged Reconstruction for Hypoplastic Left Heart Syndrome. Contemporary Results', *Annals of surgery* 224, 388.

Brown, W.M.: 1985, 'On Defining "Disease"', *Journal of Medicine and Philosophy* 10 (4), 311–328.

Bruke, G.: 1994, 'High Tech, Low Yield: Doctor's Use of Medical Innovation', *J Am Health Policy* 4 (1), 48–53.

Canguilhem, G.: 1991, *The Normal and the Pathological*. New York: Zone Books.

Cassell, E.J.: 1993, 'The Sourcer's Broom. Medicine's Rampant Technology', *Hastings Center Report* 23 (6), 32–39.

Chevenak, F.A and McCulloch, L.B.: 'Ethical Dimensions of Ultrasound Screening for Fetal Abnormalities', *Ann NY Acad Sci* 847, 185–190.

Cooper, M.W.: 1996, 'The Gastroenterologist and His Endoscope: The Embodiment of Technology and the Necessity for a Medical Ethics', *Theoretical Medicine* 17, 379–398.

Davidson, S.N.: 1995, 'Technological Cancer: Its Causes and Treatment', *Healthcare Forum J* 38 (2), 52–58.

Delkeskamp-Hayes, C. and Cutter, M.A.G.: 1993, *Science, Technology, and the Art of Medicine: European-American Dialogues*. Dordrecht: Kluwer Academic.

Ellul, J.: 1964, *The Technological Society*. New York: Alfred A. Knopf.

Evans, A.S.: 1991, 'Causation and Disease: Effect of Technology on Postulates of Causation', *The Yale Journal of Biology and Medicine* 64, 513–528.

Feinstein, A.R.: 1967, *Clinical Judgment*. Baltimore: The Williams and Wilkins Company.

Ferkiss, V.: 1969, *Technological Man*. New York: Braziller.

Fischer, E.S. and Welch, H.G.: 1999, 'Avoiding the Unintended Consequences of Growth in Medical Care', *JAMA* 281, 446–453.

Gadamer, H.G.: 1987, *Gesammelte Werke*. Tübringen: Mhor, Vol. 4.

Glover, J.: 1977, *Causing Death and Saving Lives*. Harmondsworth: Penguin.

Gorovitz, S. and MacIntyre, A.: 1976, 'Toward a Theory of Medical Fallibility', *Journal of Medicine and Philosophy* 1, 51–71.

Green, J.M.: 1990, 'Prenatal Screening and diagnosis: Some Psychological and Social Issues', *Br J Obs and Gyn* 97, 1974–1976.

Hagemo, P.S., Rasmussen, M., Bryhn, G. and Vandvik, I.H.: 1997, 'Hypoplastic Left Heart Syndrome Multiprofessional Follow-up in the Mid-term Following Palliative Procedures', *Cardiol Young* 7, 248–253.

Hanson, M.J. and Callahan, D.: 1999, *The Goals of Medicine: The Forgotten Issues in Health Care Reform*. Washington: Georgetown University Press.

Heidegger, M.: 1962, *Die Technik und die Kehre*. Stuttgart: Verlag Günther Neske.

Hellerstein, D.: 1983, 'Overdosing on Medical Technology', *Technol Rev* 86 (6), 12–17.

Helman, C.G.: 1985, 'Disease and Pseudo-disease: A Case History of Pseudo-angina', in: R.A. Hahn and A.D. Gines (eds.), *Physicians of Western Medicine. Anthropological Approaches to Theory and Practice*. Dortrecht: D. Reidel Publishing Company.

Hollander, R.D.: 1997, 'The Social Construction of Safety', in: K. Schrader-Frechett and L. Westra (eds.), *Technology and Values*. New York: Rowman & Littlefield Publishers, pp. 107–114.

Ihde, D.: 1990, *Technology and the Lifeworld: From Garden to Earth*. Bloomington: Indiana University Press.

Illich, I.: 1975, *Medical Nemesis: The Expropriation of Health*. London: Calder and Boyars.

Jennett, B.: 1986, *High Technology Medicine – Benefits and Burdens*. Oxford: Oxford University Press.

Jennett, B.: 1988, 'Variations in Surgical Practice: Welcome Diversity or Disturbing Differences', *British Journal of Surgery* 75, 630–631.

Jennett, B.: 1994, 'Medical Technology, Social and Health Care Issues', in: R. Gillon (ed.), *Principles of Health Care Ethics*. New York: John Wiley & Sons, pp. 861–872.

Jonas, H.: 1985, *Technik, Medizin und Ethik*. Frankfurt a.M: Insel Verlag.

Jonsen, A.R.: 1990, *The New Medicine and the Old Ethics*. Cambridge MA: Harward University Press.

Kass, L.R.: 1975, 'Regarding the End of Medicine and the Pursuit of Health', *Public Interest* 40, 11–42.

Kern, J.H., Hayes, C.J., Michler, R.E., Gersony, W.M. and Quagenbeur, J.M.: 1997, 'Survival and Risk Factor Analysis for the Norwood Precedure for Hypoplastic Left Heart Syndrome', *Am J Cardiol* 80 (2), 170–174.

Kevnanagh, A.M. and Broom, D.H.: 1998, 'Embodied Risk: My Body, Myself?' *Soc Sci Med* 46, 437–444.

Leape, L.L.: 1994, 'Error in Medicine', *JAMA* 23, 1851–1857.

Le Fanu, J.: 1999, *The Rise and Fall of Modern Medicine*. London: Little Brown.

Lewis, T.: 1977, 'On the Science and Technology of Medicine', in: J.H. Knowles (ed.), *Doing Better and Feeling Worse*. New York: Norton & Co, pp. 35–46.

Malone, P.S.J.: 1996, 'Antenatal Diagnosis of Renal Tract Anomalies: Has it Increased the Sum of Human Happiness?' *J R Soc Med* 89, 155–158.

McWhinney, I.R.: 1997, *A Textbook for Family Medicine*. Oxford: Oxford University Press.

Mitcham, C.: 1994, *Thinking Through Technology. The Path Between Engineering and Philosophy*. Chicago: The University of Chicago Press.

Mitcham, C.: 1995, 'Philosophy of Technology', in: W.T. Reich (ed.), *Encyclopedia of Bioethics*. New York: MacMillan, pp. 2477–1484.

Monti, J.E., Wood, D.P., Pontes, E., Boyett, J.M. and Levin H.S.: 1989, 'Adenocarcinoma of the Prostate in Cytoprostatectomy Specimens Removed from Bladder Cancer', *Cancer* 63, 531–538.

Moss, R.W.: 1991, *The Cancer Industry: The Classic Exposé on the Cancer Establishment*. New York: Paragon House.

Muraskas, J., Marshall, P.A., Tomich, P., Myers, T.F., Gianopoulos, J.G. and Thomasma, D.C.: 1999, 'Neonatal Viability in the 1990s: Held Hostage by Technology', *Camb Q Healthc Ethics* 8 (2), 160–170.

Paul, N.: 1998, 'Incurable Suffering from the "Hiatus Theoreticus"? Some Epistemological Problems in Modern Medicine and the Clinical Relevance of Philosophy of Medicine', *Theoretical Medicine and Bioethics* 19, 229–248.

Payer, L.: 1992, *Disease Mongers: How Doctors, Drug Companies, and Insurers Are Making You Feel Sick*. New York: John Wiley & Sons.

Pellegrino, E.D.: 1979, 'Medicine, Science, Art: An Old Controversy Revisited', *Man Med* 4 (1), 43–52.

Pickering, W.G.: 1996, 'Does Medical Treatment Mean Patient Benefit?' *Lancet* 347 (8998), 379–380.

Ponder, B.: 1997, 'Genetic Testing for Cancer Risk', *Science* 278, 1050–1054.

Reiser, S.J.: 1977, 'Therapeutic Choice and Moral Doubt in a Technological Age', in: J.H. Knowles (ed.), *Doing Better and Feeling Worse*. New York: Norton & Co, pp. 47–56.

Reiser, S.J.: 1978, *Medicine and the Reign of Technology*. New York: Cambridge University Press.

Robinson, D. and Bevan, E.A.: 1993, 'Defining Normailty – Art or Science?' *Methods Inf Med* 32 (3), 225–228.

Rothman D.J.: 1991, *Strangers at the Bedside. History of How Law and Bioethics Transformed Medical Decision Making*. New York: Basic Books.

Rothman, D.J.: 1997, *Beginnings Count: The Technological Imperative in American Health Care*. New York: Oxford University Press.

Räikkä, J.: 1996, 'The Social Concept of Disease', *Theoretical Medicine* 17 (4), 353–361

Schneidermann, L.J. and Jecker, N.S.: 1995, *Wrong Medicine: Doctors, Patients and Futile Treatment*. Baltimore: Johns Hopkins.

Schrader-Frechett, K. and Westra, L.: 1997, *Technology and Values*. New York: Rowman & Littlefield Publishers.

Sharpe, V.A. and Faden, A.I.: 1998, *Medical Harm. Historical, Conceptual, and Ethical Dimensions of Iatrogenic Illness*. Cambridge: Cambridge University Press.

Skrabanek, P.: 1994, *The Death of Humane Medicine and the Rise of Coercive Healthism*. Suffolk: The Social Affairs Unit.

Smith, M.R. and Marx, L. (eds.): 1994, *Does Technology Drive History? The Dilemma of Technological Determinism*. Cambridge: MIT Press.

Stewart-Brown, S. and Farmer, A.: 1997, 'Screening Could Seriously Damage your Health. Decisions to Screen Must Take Account of the Social and Psychological Costs', *BMJ* 314, 533.

Sundström, P.: 1998, 'Interpreting the Notion that Technology is Value-Natural', *Medicine, Health Care and Philosophy* 1, 41–45.

Tatum, J.S.: 1997, 'The Political Construction of Technology: A Call for Constructive Technology Assessment', in: K. Schrader-Frechett and L. Westra (eds.), *Technology and Values*. New York: Rowman & Littlefield Publishers, pp.

Temkin, O.: 1963, 'The Scientific Approach to Disease: Specific Entity and Individual Sickness', in: A.C. Crombie (ed.), *Scientific Change: Historical Studies in the Intellectual, Social and Technical Conditions for Scientific Discovery and Technical Invention from Antiquity to the Present*. New York: Basic Books, pp. 629–647.

Temkin, O.: 1977, *The Double Face of Janus*. Baltimore: Johns Hopkins University Press.

Thomas, L.: 1977, 'On the Science and Technology of Medicine', in: J.H. Knowles (ed.), *Doing Better and Feeling Worse: Health in the United states*. New York: W.W. Norton.

Toulmin S.: 1986, 'How Medicine Saved the Life of Ethics', in: DeMarco, R.M. Fox (eds.), *New Directions in Ethics: The Challenge of Applied Ethics*. New York: Routledge and Keagan Paul, pp. 265–281.

Twaddle, A.: 1993, 'Disease, Illness and Sickness Revisited', in: A. Twaddle and L. Nordenfelt (eds.), *Disease, Illness and Sickness: Three Concepts in the Theory of Health, Studies in Health and Society 18*, Linköping University, pp. 1–18.

Tijmstra, T.: 1989, 'The Imperative Character of Medical Technology and the Meaning of "Anticipated Decision Regret" ', *Int J of Technology Assessment in Health Care* 5, 207–213.

Veatch, R.M.: 1985, 'Against Virtue: A Deontological Critique of Virtue Theory in Medical Ethics', in: E.E. Shelp (ed.), *Virtue and Medicine*. Dordrecht: Reidel Publishing Company, pp. 329–345.

Vos, R.: 1991, *Drugs Looking for Diseases: Innovative Drug Research and the Development of the Beta Blockers and Calcium Antagonists*. Dordrecht: Kluwer Academic Pbulishers.

Wahl, P.W., Savage, P.J., Psaty, B.M., Orchard, T.J., Robbins, J.A. and Tracy, R.P.: 1998, 'Diabetes in Older Adults: Comparison of 1997 American Diabetes Association Classification of Diabetes Mellitus with 1985 WHO Classification', *Lancet* 352 (9133), 1012–1015.

Whittle, M.: 1997, 'Ultrasonographic "Soft Markers" of Fetal Ultrasound. Detecting Them may do More Harm Than Good' (Editorial). *BMJ* 314, 918.

Winner, L.: 1977, *Autonomous Technology*. Cambridge MA: MIT Press.

Part III
AI in healthcare:

Framing challenges,

current trends

and

future possibilities

OPEN ACCESS

Medicine and the rise of the robots: a qualitative review of recent advances of artificial intelligence in health

Erwin Loh

Correspondence to
Professor Erwin Loh, Monash Centre for Health Research and Implementation, Monash University, Clayton, VIC 3168, Victoria, Australia;
erwin.loh@monash.edu

Received 14 February 2018
Revised 1 May 2018
Accepted 11 May 2018
Published Online First 1 June 2018

ABSTRACT

Artificial intelligence (AI) has the potential to significantly transform the role of the doctor and revolutionise the practice of medicine. This qualitative review paper summarises the past 12 months of health research in AI, across different medical specialties, and discusses the current strengths as well as challenges, relating to this emerging technology. Doctors, especially those in leadership roles, need to be aware of how quickly AI is advancing in health, so that they are ready to lead the change required for its adoption by the health system. Key points: 'AI has now been shown to be as effective as humans in the diagnosis of various medical conditions, and in some cases, more effective.' When it comes to predicting suicide attempts, recent research suggest AI is better than human beings. 'AI's current strength is in its ability to learn from a large dataset and recognise patterns that can be used to diagnose conditions, putting it in direct competition with medical specialties that are involved in diagnostic tests that involve pattern recognition, such as pathology and radiology'. The current challenges in AI include legal liability and attribution of negligence when errors occur, and the ethical issues relating to patient choices. 'AI systems can also be developed with, or learn, biases, that will need to be identified and mitigated'. As doctors and health leaders, we need to start preparing the profession to be supported by, partnered with, and, in future, potentially be replaced by, AI and advanced robotics systems.

INTRODUCTION

Artificial intelligence (AI) has been defined by Alan Turing, the founding father of AI, as 'the science and engineering of making intelligent machines, especially intelligent computer programs'.[1] AI in health uses algorithms and software to approximate the cognition undertaken by human clinicians in the analysis of complex medical data. AI research has been divided into subfields, based on goals such as machine learning or deep learning, and tools such as neural networks, a subset of machine learning.[2] AI has the potential to significantly transform the role of the doctor and revolutionise the practice of medicine, and it is important for all doctors, in particular those in positions of leadership within the health system, to anticipate the potential changes, forecast their impact and plan strategically for the medium to long term.

The impact of automation and robotics have been felt by blue-collar jobs for a while. A recent working paper by the National Bureau of Economic Research found that the arrival of one new industrial robot in a local labour market coincides with an employment drop of 5.6 workers.[2] Last year alone, there have been news reports of apple-picking robots,[3] burger-flipping robots[4] and a barista robot that makes you coffee.[5] *Nature* even ran an editorial on sex robots.[6]

There is a false sense of security in assuming that automation will only impact blue-collar type work that requires more manual, repetitive actions and less intellectual input. PwC released a report based on a survey of 2500 US consumers and business leaders, which predicts that AI will continue to make in-roads into white collar industries.[7] A large stockbroking firm ran a trial in Europe of its new AI program this year that showed it was much more efficient than traditional methods of buying and selling shares.[8] A Japanese insurance firm replaced 34 employees with an AI system, which it believes will increase productivity by 30% and see a return on its investment in less than 2 years.[9] The *Washington Post* used an AI reporter to publish 850 articles in the past year.[10]

Not even the jobs of computer programmers, the creators of the code for AI, are safe. Microsoft and Cambridge built an AI capable of writing code that would solve simple math problems.[11] Lawyers are not exempt either. Late last year, an AI was able to predict the judicial decisions of the European Court of Human Rights with 79% accuracy.[12]

Compared with other industries like hospitality or airlines, health has been a relative slow adopter of electronic systems, such as electronic health record (EHR) systems, which have only recently become mainstream.[13] Similarly, although AI is now embedded in many forms of technologies such as smartphones and software, its use in the front-line of clinical practice remains limited. Nevertheless, research in this area continues to grow exponentially.

QUALITATIVE REVIEW METHODOLOGY

This paper summarises the past 12 months of health research in AI, across different medical specialties, and discusses the current strengths and weaknesses, as well as challenges, relating to this emerging technology. The author notes that much progress has been made by AI developments in health over the past two to three decades and has focused on the past 12 months because of some of the exponential gains made, mainly due to improvements in computer hardware technologies. The author has specifically restricted his review to recent research in AI published in high-ranking peer-reviewed medical

Check for updates

To cite: Loh E. *BMJ Leader* 2018;**2**:59–63.

leader: first published as 10.1136/leader-2018-000071 on 1 June 2018. Downloaded from http://bmjleader.bmj.com/ on November 6, 2019 by guest. Protected by copyright.

journals. The selection criteria involved keywords relating to artificial intelligence, machine learning, deep learning and algorithms relating to medical diagnosis, planning and treatment.

This qualitative review is not intended to be a systematic review, and the author has restricted the research to AI research that will likely to have the most impact to clinical practice, a judgement that is subjective to the author's own experience and expertise as a specialist medical administrator in both academia and practice. The time period of around 12 months is because the exponential growth and improvements in AI technology means that any data presented that are older may no longer be applicable.

The focus of the review is to provide a high-level update of recent AI research in health to ensure that medical practitioners, especially those in leadership roles, are made aware of how quickly AI is advancing in health, so that they are made ready to lead the change required for its adoption by the health system.

FINDINGS

AI in medical diagnosis

AI has now been shown to be effective in the accurate diagnosis of various medical conditions. For example, in ophthalmology, an AI-based grading algorithm was used to screen fundus photographs obtained from diabetic patients and identify, with high reliability (94% and 98% sensitivity and specificity), to determine cases that should be referred to an ophthalmologist for further evaluation and treatment.[14] In another study, researchers showed that an AI agent, using deep learning and neural networks, accurately diagnosed and provided treatment decisions for congenital cataracts in a multihospital clinical trial, performing just as well as individual ophthalmologists.[15]

In relation to skin cancer, researchers trained a neural network using a dataset of 129 450 clinical images and tested its performance against 21 board-certified dermatologists on biopsy-proven clinical images. The neural network achieved performance on par with all tested experts, demonstrating that an AI was capable of classifying skin cancer with a level of competence comparable with dermatologists.[16] In another study using routine clinical data of over 350 000 patients, machine learning significantly improved accuracy of cardiovascular risk prediction, correctly predicting 355 (additional 7.6%) more patients who developed cardiovascular disease compared with the established algorithm.[17]

Clinical neuroscience has also benefited from AI. A deep-learning algorithm used MRI of the brain of individuals 6 to 12 months old to predict the diagnosis of autism in individual high-risk children at 24 months, with a positive predictive value of 81%.[18] Similarly, in another study, a machine learning method designed to assess the progression to dementia within 24 months, based on a single amyloid PET scan, obtained an accuracy of 84%, outperforming the existing algorithms using the same biomarker measures and previous studies using multiple biomarker modalities.[19]

AI in psychiatry

AI may be good at diagnosing physical illness, but what about its use in psychological medicine and psychiatry? The emerging literature has also shown that AI is proving to be useful in these clinical areas. For example, researchers built a predictive model based on machine learning using whole-brain functional magnetic resonance imaging (fMRI) to achieve 74% accuracy in identifying patients with more severe negative and positive symptoms in schizophrenia, suggesting the use of brain imaging to predict the disease and its symptom severity.[20] In another study, researchers demonstrated that a linguistic machine learning system, using fMRI and proton magnetic resonance spectroscopy (^{1}H-MRS) inputs, showed nearly perfect classification accuracy and was able to predict lithium response in bipolar patients with at least 88% accuracy in training and 80% accuracy in validation, allowing psychiatrists the ability to predict lithium response and avoid unnecessary treatment.[21]

It is one thing for AI to be able to recognise patterns on images from radiology and pathology tests. Can AI be as good as psychiatrists when it comes to predicting mental health conditions that do not have a clear biomarker? A landmark paper of a meta-analysis of 365 studies spanning 50 years published by the American Psychological Association found that prediction of suicide was only slightly better than chance for all outcomes, and that this predictive ability has not improved across 50 years of research, leading the authors to suggest the need for a shift in focus from risk factors to machine learning-based risk algorithms.[22]

Researchers at the Vanderbilt University Medical Centre created machine-learning algorithms that achieved 80%–90% accuracy when predicting whether someone will attempt suicide within the next 2 years, and 92% accuracy in predicting whether someone will attempt suicide within the next week, by applying machine learning to patients' EHRs. In other words, when it comes to predicting suicide attempts, AI appears to be better than human beings, although the clinical applicability in the real world remains unproven.[23] In another study, researchers used machine-learning algorithms to identify individuals at risk of suicide with high (91%) accuracy, based on their altered fMRI neural signatures of death-related and life-related concepts.[24] These developments in AI are now being applied. Facebook is one of several companies exploring ways to use AI algorithms to predict suicide based on mining social media.[25]

AI in treatment

So, we have established that AI can be helpful in predicting mental health conditions, but can AI also be helpful in the provision of psychological treatments? Researchers found that soldiers are more likely to open up about post-traumatic stress when interviewed by a computer-generated automated virtual interviewer, and such virtual interviewers were found to be superior to human ones in obtaining more psychological symptoms from veterans.[26]

What about robot surgeons? Robotic surgical devices already exist, but they still require human control—is AI able to perform autonomous surgery without human input? In a robotic surgery breakthrough in 2016, a smart surgical robot stitched up a pig's small intestines completely on its own and was able to do a better job on the operation than human surgeons who were given the same task.[27] What is even more impressive is that late last year, a robot dentist in China was able to carry out the world's first successful autonomous implant surgery by fitting two new teeth into a woman's mouth without any human intervention.[28]

AI's current strengths

So, based on the available evidence, what is AI good at today? It is clear that AI's current strength is in its ability to learn from a large dataset and recognise patterns that can be used to diagnose conditions. This puts AI in direct competition with medical specialties that are involved in diagnostic tests that involve pattern recognition, and the two obvious ones are pathology and radiology.

leader: first published as 10.1136/leader-2018-000071 on 1 June 2018. Downloaded from http://bmjleader.bmj.com/ on November 6, 2019 by guest. Protected by copyright.

An editorial on recent studies point to the future of computational pathology, suggesting that computers will increasingly become integrated into the pathology workflow when they can improve accuracy in answering questions that are difficult for pathologists.[29] However, Google researchers used an AI in a study to identify malignant tumours in breast cancer images with an 89% accuracy rate, compared with 73% achieved by a human pathologist.[30] In another study, deep learning algorithms achieved better diagnostic performance than a panel of 11 pathologists, in a simulated time-constrained diagnostic setting, in detecting lymph node metastases in tissue sections of women with breast cancer.[31]

Similarly, radiologists are grappling with the potentially disruptive applications of machine learning to image analysis in their specialty, but remain as a profession optimistic that AI will be able to provide opportunities for radiologists to augment and improve the quality of care they provide to their patients.[32] However, AI systems continue to improve in their diagnostic and predictive capabilities in radiology. For example, a machine-learning model, using three-dimensional cardiac motion on cardiac MRI, was able to predict survival outcome independent of conventional risk factors in patients with newly diagnosed pulmonary hypertension.[33] It is also interesting to note that the first United States Food and Drugs Administration approval for an AI application in a clinical setting is for a deep learning platform in radiology, to help doctors diagnose heart problems.[34]

Can AI completely replace the role of a doctor?

AI may be as good as, or even better than, humans when it comes to formulating diagnoses based on recognising patterns on images, but is AI ready to take over the complete role of a fully trained medical practitioner? So far, the answer appears to be—not yet. In the first direct comparison of diagnostic accuracy, physicians were found to vastly outperform computer algorithms in diagnostic accuracy (84.3% vs 51.2% correct diagnosis in the top three listed).[35] Bear in mind that this study compared doctors with relatively simple symptom checker applications.

In a more recent study, Watson, IBM's AI platform, took just 10 min to analyse a genome of a patient with brain cancer and suggest a treatment plan, compared with human experts who took 160 hours to make a comparable plan.[36] In another study, Watson found cancer treatments that oncologists overlooked, by discovering 'potential therapeutic options' for 323 additional patients after analysing 'large volumes of data', including past studies, databases and genetic information.[37] It should be noted that these superior performances in the theoretical setting has not translated well into real-world clinical practice, based on recent reports of poor clinician adoption at a major American cancer centre.[38]

As such, it would seem that AI systems may be better than human doctors in coming with diagnoses or management plans, if they are provided with sufficiently large amounts of data that are beyond what humans can manually analyse.

DISCUSSION

Challenges of AI in health

It is clear from the qualitative literature review that AI in health has progressed remarkably, even within the span of 12 months looked at. It is likely that much of this recent progress is due to the increasing presence of large training data sets and improvements in computer hardware, in the form of memory and computational capacity. However, there are some challenges that need to be considered as AI usage increases in healthcare. One of the concerns that has been raised is the issue of legal liability. If a medical error occurs, who is to be held liable? A robot surgeon is not a legal entity, so should the patient sue the owner, the programmer, the manufacturer or someone else? Could an AI ever be subject to criminal liability? These AI dilemmas are not unique to health—for example, there have already been a few high-profile self-driving car accidents, some resulting in fatalities. These are some of the issues that legal experts have been grappling with that are still unresolved.[39]

The other issue to consider is the potential for AI to greatly reduce the number of medical errors and misdiagnoses, and therefore reduce medicolegal claims. What happens when the ability of AI surpasses that of the average doctor? If a doctor relies on the recommendation of an AI tool, which ends up being wrong, is it still the negligence of the doctor if that tool has already been proven to be more reliable that the average doctor? An argument has been put forth, although under the US legal system, to suggest that a by-product of an increased use of AI in health is that doctors will practise less defensive medicine, by ordering less unnecessary tests, because they will be relying on the recommendations of AI systems that are better diagnosticians than they are.[40] In fact, there may come a day that it would be considered negligent for a doctor *not* to consider the recommendation of a health AI system if that becomes the standard of care.

There is also the matter of morality and ethics with AI. The best way to illustrate this issue is by describing the classic 'trolley problem'—if you are in a trolley that is going down a track that is about to hit five workers, and you can redirect the trolley by turning it onto another track but there is one worker on it, is it morally permissible to turn the trolley to spare the lives of five workers by killing the single worker?[41] This dilemma is particularly pertinent to self-driving cars, as that scenario could realistically actually happen in real life—what should the self-driving car in the event of an accident do in an attempt to reduce the number of injured humans? Should the self-driving car prioritise the passengers over the pedestrians? Who gets to make these decisions? The programmer or the passenger?

Researchers have attempted to resolve this issue by suggesting that self-driving cars be equipped with what they call an 'Ethical Knob', a device enabling passengers to ethically customise their autonomous vehicles to choose between different settings corresponding to different moral approaches or principles. In this way, the AI in self-driving cars would be entrusted with implementing users' ethical choices, while manufacturers/programmers would be tasked with enabling the user's choice.[42] Similarly, an AI in healthcare can be provided guidance as to the moral wishes of the patient—for example, does the patient want to maximise length of life or the quality of life?

This brings us to another real issue with AI—inherent bias. AI systems can be inadvertently programmed to have bias because of the biases of the programmers or, with the development of self-learning algorithms, actually learn to be biased based on the data it is learning from. In addition, AI systems find it more difficult to generalise findings from a narrower dataset, with minor differences from a training set potentially making larger-than-intended impact on a prospective set of data, creating potential bias. A recent study demonstrated that AI can learn to have racist or sexist biases, based on word associations that are part of data it was learning from, sourced from the internet that reflected humanity's own cultural and historical biases.[43] Strategies to minimise and mitigate such biases will need to be in place as adoption of AI by health increases.

The last issue that needs to be considered relates to how AI uses data. In the past, EHR systems used to require that data be

leader: first published as 10.1136/leader-2018-000071 on 1 June 2018. Downloaded from http://bmjleader.bmj.com/ on November 6, 2019 by guest. Protected by copyright.

properly entered into the correct categories for the right queries to be made to extract useful information. However, the advent of fuzzy logic, a form of AI, now allows for free-text unstructured text to be queried and categorised in real time to provide meaningful information.[44] The quality of the information extracted is still dependent on the accuracy of the data being entered, as patient-reported outcome measures may still be unreliable.[45] In addition, sophisticated AI systems can link disparate health data from separate databases together to form connections that may otherwise be missed.

As such, AI is now being applied to the large health data repositories because of the amount of free-text stored and also because AI, through machine learning, needs access to vast amounts of data. However, the issue of data ownership and privacy needs to be considered. A relevant case study is the recent finding by the UK's Information Commissioner that a National Health Service trust breached privacy laws by sharing patient data with Google for Google's DeepMind Streams app.[46] Although this app did not directly use AI, the alleged data breach demonstrates that need for the development of a data governance framework that takes into account data ownership, privacy principles, patient consent and data security.[47] Current privacy laws may need to be reviewed to ensure they are relevant even as social media and other large technologies like Google start using AI to commercialise the big data they have collected from their millions of users.

Future of AI

There is no turning back from the rise of AI in all aspects of our lives. AI already resides in the smartphones that a lot of us own, in the form of smart digital assistants. But AI has progressed beyond helpful chatbots. For example, Google's AI group, Deepmind, unveiled AlphaGo, an AI that took just 3 days to master the ancient Chinese board game of Go with no human input, as reported in *Nature*.[48] This version of AI was able to win against its previous version (that famously beat the world champion in Go previously) 100 games to 0. More recently, AlphaZero, another AI from Google, learnt the rules of chess in 4 hours by playing against itself 44 million times and went on to beat Stockfish, a well-established chess program.[49]

AI researchers are already developing AI algorithms that are able to learn, grow and mature like human beings do, through self-reflection[50] and experiencing the world firsthand.[51] AI can currently analyse large amounts of data much faster than humans can using today's hardware. However, quantum computers, which may outperform the classical computers we have today by many factors, are already in development and only a few years away.[52] In addition, scientists have made a pioneering breakthrough by developing photonic computer chips—that use light rather than electricity—that imitate the way the brain's synapses operate, which means that computers may be able to process data at the speed of light in the near future, compared with human nerve conduction speed that is slower than electricity as it is.[53]

With dramatic improvements in computer software and hardware coming online, and increasing access to large datasets that are increasingly being linked together, it is no wonder that Ray Kurzweil, a Google AI expert and well-known futurist, believes that AI will surpass the brainpower of a human being by 2023 and reach what he terms 'singularity' in 2045, which is when AI will surpass the brainpower equivalent to that of all human beings combined.[54]

Implications for medical leaders

Those of us who are medical leaders in healthcare, in particular, in the public health system, know that the health system is traditionally risk averse and tends to be a slower adopter of new technologies. Nevertheless, it is essential that medical leaders like us are aware of the potential impacts that new health technologies will have on the current and future health system.

As such systems are introduced into our health services, medical leaders need to ensure that there are strong and robust governance structures in place to ensure that there is appropriate review of these new technologies prior to implementation, in terms of their safety, cost-effectiveness and that staff are credentialled to use the new technologies. A data governance framework will also be required to oversee how data are managed internally, the data standards and quality expected, how data are received, how data are secured and how data are shared externally to different stakeholders, in compliance with relevant laws and regulations. An appropriate training regime should also be implemented to ensure that staff are aware of their ethical and legal responsibilities when it comes to data management, especially as it relates to the use of social media.

Medical leaders will also need to constantly scan the horizon for future developments in the field of AI, and consider future risks and opportunities, in order to plan accordingly. AI and automation will have an impact of the health workforce, and workforce planning will need to take this issue into account. The opportunities offered by AI to improve the care of patients need to be taken into account when new IT systems are introduced, in particular, where AI can assist in interrogating large amounts of health data, which may be unstructured or separated into different silos.

Medical leaders should also be aware that AI systems are not just relevant for clinical care—AI systems are increasingly being applied in the management setting. AI can be used to support, and potentially replace, the role of managers, including in health, in financial management, priority setting, resource allocation and workforce management. We will need to consider how AI can support us in our roles, now and into the future.

Lastly, medical leaders will need to be change agents and lead the change as AI transforms the healthcare system in the coming years. We will need to ensure that the patient experience and needs are always prioritised, and that compassion and kindness are not replaced by efficiencies and metrics. As leaders of clinicians, we will need to manage the anxiety of the clinical workforce through potential uncertain times, by refocussing any changes on improving patient care. Ultimately, medical leaders are still doctors, and our duty of care is to our patients.

CONCLUSION

It is evident from this qualitative review of recent evidence that AI research in health continue to progress, and that AI is proving to be effective in most aspects of medicine, including diagnosis, planning and even treatment. As a profession, we need to have a mature discussion and debate about the legal, ethical and moral challenges of AI in health, and mitigate any potential bias that such systems may inherit from their makers.

Regardless of whether the AI singularity comes to pass or not, AI in health will continue to improve, and these improvements appear to be accelerating. There are clear challenges for the adoption of AI in health for health services, organisations and governments, and a need to develop a policy framework around this issue. As doctors and health leaders, we need to start preparing the profession to be supported by, partnered with, and, in future,

leader: first published as 10.1136/leader-2018-000071 on 1 June 2018. Downloaded from http://bmjleader.bmj.com/ on November 6, 2019 by guest. Protected by copyright.

potentially be replaced by, AI and advanced robotics systems. We have an opportunity now to literally shape the development of humanity's future autonomous health providers, and we should be leaders in this space rather than passive observers.

Contributors EL planned, conducted and submitted the study.

Funding The author has not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/

© Article author(s) (or their employer(s) unless otherwise stated in the text of the article) 2018. All rights reserved. No commercial use is permitted unless otherwise expressly granted.

REFERENCES

1 Turing AM. I.—Computing machinery and intelligence. *Mind* 1950;LIX:433–60.
2 Acemoglu D, Restrepo P. *Robots and jobs: evidence from US labor markets*, 2017. NBER Working Paper No. 23285.
3 Scientific American. What an Apple-Picking Robot Means for the Future of Farm Workers. [Online]. 2017 https://www.scientificamerican.com/article/what-an-apple-picking-robot-means-for-the-future-of-farm-workers/ (cited 1 Nov 2017).
4 Inc. Meet Flippy the Burger-Flipping Robot (He's Hungry to Steal Your Fast-Food Job). [Online]. 2017 https://www.inc.com/peter-economy/meet-flippy-the-burger-flipping-robot-and-hes-hung.html (cited 1 Nov 2017).
5 Wired. This Robot Barista Makes a Dang Good Latte. [Online]. 2017 https://www.wired.com/2017/01/cafe-x-robot-barista/ (cited 1 Nov 2017).
6 Nature. Let's talk about sex robots. *Nature* 2017;547:138.
7 PwC. *Bot.Me: A Revolutionary Partnership*, 2017. Consumer Intelligence Series.
8 Financial Times. JPMorgan develops robot to execute trades. 2017 https://www.ft.com/content/16b8ffb6-7161-11e7-aca6-c6bd07df1a3c (cited 1 Nov 2017).
9 Guardian. Japanese company replaces office workers with artificial intelligence. [Online]. 2017 https://www.theguardian.com/technology/2017/jan/05/japanese-company-replaces-office-workers-artificial-intelligence-ai-fukoku-mutual-life-insurance (cited 1 Nov 2017).
10 Digiday. The Washington Post's robot reporter has published 850 articles in the past year. [Online]. 2017 https://digiday.com/media/washington-posts-robot-reporter-published-500-articles-last-year/ (cited 1 Nov 2017).
11 Balog M, Gaunt A, Brockschmidt M, *et al*. DeepCoder: learning to write programs. *ICLR* 2017.
12 Aletras N, Tsarapatsanis D, Preoțiuc-Pietro D, *et al*. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Computer Science* 2017;2:e92.
13 Palabindala V, Pamarthy A, Jonnalagadda NR. Adoption of electronic health records and barriers. *J Community Hosp Intern Med Perspect* 2016;6:32643.
14 Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology* 2017;124:962–9.
15 Long E, Lin H, Liu Z, *et al*. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. *Nat Biomed Eng* 2017;1:0024.
16 Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115–8.
17 Weng SF, Reps J, Kai J, *et al*. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One* 2017;12:e0174944.
18 Hazlett HC, Gu H, Munsell BC, *et al*. Early brain development in infants at high risk for autism spectrum disorder. *Nature* 2017;542:348–51.
19 Mathotaarachchi S, Pascoal TA, Shin M, *et al*. Identifying incipient dementia individuals using machine learning and amyloid imaging. *Neurobiol Aging* 2017;59:80–90.
20 Gheiratmand M, Rish I, Cecchi GA, *et al*. Learning stable and predictive network-based patterns of schizophrenia and its clinical symptoms. *NPJ Schizophr* 2017;3:22.
21 Fleck DE, Ernest N, Adler CM, *et al*. Prediction of lithium response in first-episode mania using the LITHium Intelligent Agent (LITHIA): pilot data and proof-of-concept. *Bipolar Disord* 2017;19:259–72.
22 Franklin JC, Ribeiro JD, Fox KR, *et al*. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychol Bull* 2017;143:187–232.
23 Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. *Clin Psychol Sci* 2017;5:457–69.
24 Just MA, Pan L, Cherkassky VL, *et al*. Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nat Hum Behav* 2017;1:911–9.
25 Reardon S. AI algorithms to prevent suicide gain traction. *Nature* 2017;64. ISSN 1476-4687.
26 Lucas GM, Rizzo A, Gratch J, *et al*. Reporting mental health symptoms: breaking down barriers to care with virtual human interviewers. *Frontiers in Robotics and AI* 2017;4:51.
27 Shademan A, Decker RS, Opfermann JD, *et al*. Supervised autonomous robotic soft tissue surgery. *Sci Transl Med* 2016;8:337ra64.
28 South China Morning Post. Chinese robot dentist is first to fit implants in patient's mouth without any human involvement. [Online]. 2017 http://www.scmp.com/news/china/article/2112197/chinese-robot-dentist-first-fit-implants-patients-mouth-without-any-human (cited 10 Jan 2018).
29 Granter SR, Beck AH, Papke DJ. AlphaGo, Deep Learning, and the Future of the Human Microscopist. *Arch Pathol Lab Med* 2017;141:619–21.
30 Liu Y, Gadepalli K, Norouzi M, *et al*. Detecting cancer metastases on gigapixel pathology images. *arXiv* 2017:1703.
31 Ehteshami Bejnordi B, Veta M, Johannes van Diest P, *et al*. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017;318:2199–210.
32 Kruskal JB, Berkowitz S, Geis JR, *et al*. Big Data and Machine Learning—Strategies for Driving This Bus: A Summary of the 2016 Intersociety Summer Conference. *J Am Coll Radiol* 2017;14:811–7.
33 Dawes TJW, de Marvao A, Shi W, *et al*. Machine learning of three-dimensional right ventricular motion enables outcome prediction in pulmonary hypertension: a cardiac MR imaging study. *Radiology* 2017;283:381–90.
34 Forbes. First FDA Approval For Clinical Cloud-Based Deep Learning In Healthcare. [Online]. 2017 https://www.forbes.com/sites/bernardmarr/2017/01/20/first-fda-approval-for-clinical-cloud-based-deep-learning-in-healthcare/ (cited 11 January 2018).
35 Semigran HL, Levine DM, Nundy S, *et al*. Comparison of physician and computer diagnostic accuracy. *JAMA Intern Med* 2016;176:1860–1.
36 Wrzeszczynski KO, Frank MO, Koyama T, *et al*. Comparing sequencing assays and human-machine analyses in actionable genomics for glioblastoma. *Neurol Genet* 2017;3:e164.
37 Patel NM, Michelini VV, Snell JM, *et al*. Enhancing next-generation sequencing-guided cancer care through cognitive computing. *Oncologist* 2018;23:179–85.
38 Stat News. IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close. [Online]. 2017 https://www.statnews.com/2017/09/05/watson-ibm-cancer/ (cited 3 Apr 2018).
39 Kingston J. Artificial Intelligence and Legal Liability. *Research and Development in Intelligent Systems XXXIII*. Cham, Switzerland: Springer, 2016.
40 Thomas S. *Artificial Intelligence, Medical Malpractice, and the End of Defensive Medicine*, 2017.
41 Thomson JJ. The trolley problem. *Yale Law J* 1985;94:1395–415.
42 Contissa G, Lagioia F, Sartor G. The Ethical Knob: ethically-customisable automated vehicles and the law. *Artificial Intelligence and Law* 2017;25:365–78.
43 Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. *Science* 2017;356:183–6.
44 Yager RR. Fuzzy logics and artificial intelligence. *Fuzzy Sets Syst* 1997;90:193–8.
45 Frost MH, Reeve BB, Liepa AM, *et al*. What is sufficient evidence for the reliability and validity of patient-reported outcome measures? *Value Health* 2007;10:S94–105.
46 BBC News. Google DeepMind NHS app test broke UK privacy law. [Online]. 2017 http://www.bbc.com/news/technology-40483202 (cited 3 Apr 2018).
47 Commissioners 3ICoDPaP. *Artificial Intelligence, Robotics, Privacy and Data Protection*. Marrakech: European Data Protection Supervisor, 2016:3. In Room document.
48 Silver D, Schrittwieser J, Simonyan K, *et al*. Mastering the game of Go without human knowledge. *Nature* 2017;550:354–9.
49 Silver D, Hubert T, Schrittwieser J, *et al*. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv* 2017.
50 Jaderberg M, Mnih V, Czarnecki WM, *et al*. Reinforcement learning with unsupervised auxiliary tasks. *arXiv* 2016.
51 Denil M, Agrawal P, Kulkarni T, *et al*. Learning to perform physics experiments via deep reinforcement learning. *arXiv* 2017:1611.01843.
52 Zhang J, Pagano G, Hess PW, *et al*. Observation of a many-body dynamical phase transition with a 53-qubit quantum simulator. *Nature* 2017;551:601–4.
53 Cheng Z, Ríos C, Pernice WHP, *et al*. On-chip photonic synapse. *Sci Adv* 2017;3:e1700160.
54 Kurzweil R. *The Singularity Is Near: When Humans Transcend Biology*. New York: Penguin, 2006.

leader: first published as 10.1136/leader-2018-000071 on 1 June 2018. Downloaded from http://bmjleader.bmj.com/ on November 6, 2019 by guest. Protected by copyright.

Framing the challenges of artificial intelligence in medicine

Kun-Hsing Yu, Isaac S Kohane

Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA

Correspondence to
Dr Kun-Hsing Yu and Professor Isaac S Kohane, Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA;
Kun-Hsing_Yu@hms.harvard.edu,
isaac_kohane@harvard.edu

Received 4 July 2018
Revised 20 August 2018
Accepted 26 August 2018

Check for updates

© Author(s) (or their employer(s)) 2018. No commercial re-use. See rights and permissions. Published by BMJ.

To cite: Yu K-H, Kohane IS. *BMJ Qual Saf* Epub ahead of print: [please include Day Month Year]. doi:10.1136/bmjqs-2018-008551

On a clear January morning in Florida, a Tesla enthusiast and network entrepreneur was driving his new Tesla Model S on US Highway 27A, returning from a family trip. He had posted dozens of widely circulated YouTube tutorial videos on his vehicle and clearly understood many of the technical details of his car. That day, he let the vehicle run autonomously on Autopilot mode for 37 min, before it crashed into the trailer of a truck turning left. The Autopilot did not identify the white side of the trailer as a potential hazard, and the driver was killed, leaving his family and his high-tech business behind.[1] This tragedy is not a metaphor for artificial intelligence (AI) applications but an example of a long-recognised challenge in AI: the Frame Problem.[2] Although rarely appreciated in the scholarly and lay descriptions of the stunning recent successes of AI in medical applications, the Frame Problem and related AI challenges will have unintended harmful effects to the care of patients if not directly addressed.

With the recent advancement in machine learning algorithms, many medical tasks previously thought to require human expertise have been replicated by AI systems at or above the level of accuracy in human experts. These important demonstrations range from evaluating fundus retinography[3] and histopathology[4] to reading chest radiographs[5] and assessment of skin lesions.[6] These studies have encompassed very large numbers of patient cases and have been extensively benchmarked against clinicians. However, all these studies are retrospective in that they involve a collection of labelled cases against which the AI systems are trained and another collection against which they are tested or validated. So far, they have not entered into routine prospective use in the clinic where the Frame Problem will manifest itself most pathologically.

The Frame Problem was first introduced by computer scientists and cognitive science pioneers McCarthy and Hayes[2] in 1969 and revolved around the difficulty in identifying and updating a set of axioms to properly describe the environment for autonomous agents. To provide a medical example, let us define a worrisome chest X-ray as being one in which a shadow or a density appears that resembles those seen in lung cancer, pneumonia or various pulmonary pathologies. As in the recent successes, we are confident that an AI program can be trained with enough well-curated cases to give accuracies greater than 90% and better than the typical or even expert radiologist.[5] What if the X-ray technician leaves on patient Jill Doe the adhesive ECG lead connectors from her recent inpatient ECG. Will the AI program classify these circular medical artefacts as one of the known chest lesions? That false positive would soon become apparent and the AI engineers would include these circular ECG leads in the training sets and that error would be eliminated. What if the ECG leads superimposed part of a real shadow of a lung nodule such that the AI program would miss it? Presumably, after a few such cases where the nodule would become clinically obvious, the AI engineers would ensure that the training sets would have enough cases of such overlap to give adequate sensitivity and specificity in these instances. What if Jill Doe, despite the technician's warning, had placed her hand with a wedding ring on her chest? If no one except the AI program looks at the image, would automated classification dismiss the ring as a non-medical artefact or would it classify it incorrectly as a lesion? If the AI program is trained to recognise such non-medical artefacts, then how will it classify a toddler—Jane Doe's—chest X-ray if she comes in with stridor and shortness of

BMJ Qual Saf: first published as 10.1136/bmjqs-2018-008551 on 5 October 2018. Downloaded from http://qualitysafety.bmj.com/ on 7 October 2018 by guest. Protected by copyright.

breath? Specifically, with a ring visible in the image because Jane Doe had aspirated her mother's ring? In each of these cases, a reasonable response is that the program could be further trained and adjusted, or a human overseeing the process could use their common sense and experience to intervene. If the former, then the central question is how rapidly will training or adjusting reach acceptable performance? If the latter, how efficiently can a human oversee these programs with consistent vigilance while preserving the cost and accuracy benefits? Fortunately, these are empirically answerable questions. Unfortunately, they can only be answered through prospective trials if AI programs in medical care are to have the public and professional trust required for their full impact.

Trust in medical technology is closely related to its anticipated utility. Disruptions of the current clinical workflow will inevitably face inertia, and if the perception emerges that there are untoward consequences of a new technology then the barrier to any similar technologies will become next to insurmountable. The multidecade hiatus in gene therapy[7 8] and the retardant effect of the Theranos debacle on fingerprick blood diagnostics[9] are two more recent examples. This issue becomes especially complicated if the technology is complex or if the implementation details are proprietary, such that the general public or even domain experts cannot fully evaluate its efficacy and potential hazard based on the information they received. Moreover, even in cases of overwhelming public good such as in vaccination programs, lack of full disclosure and inadequate attention to patient education and autonomy can have dramatic negative consequences for the diffusion of helpful technologies.[10] The relevance here is that the Frame Problem and related issues will inevitably cause medical errors that will draw the attention of both the public and, at least in the USA, lawsuits against parties using, deploying or developing medical AI applications. The 'black-box' nature of many modern machine learning algorithms will further exacerbate the issue. High-profile examples of harmful or inadequate performance will bring extra scrutiny on the whole field and may retard the further development of even more robust AI systems.

Furthermore, data-driven AI algorithms are not immune from the 'garbage-in-garbage-out' rule. Machine learning algorithms are designed to identify the hidden patterns of the data and generate output predictions based on what they have seen in the past.[11] As many input data sets contain artefacts or biases, the models learnt from the data carry the biases and can potentially amplify them. For instance, electronic medical records and insurance claims data sets are records of patients' clinical courses and also a tool for healthcare providers to justify specific levels of reimbursement. Consequently, optimised machine learning models in healthcare can be confounded by their training data where the reimbursement strategies driving diagnostic coding are implicit and may not reflect a more objective clinical assessment. Additionally, AI systems could perpetuate racial bias since the biases exist in historical data,[12] resulting in partiality in the seemingly 'objective' computational methods. Ongoing data-driven controversies regarding the causes of poor health outcomes among disadvantaged populations (ie, differential access vs differential biological health risks)[13] illustrate just how difficult it is to avoid confounding in the analysis of observational health data. Paying a lot more attention to data quality and provenance, an expensive proposition, will go a long way to foster 'patient trust' in medical AI systems, and to avoid unethical medical performance, even if only by negligence.

Last but not least, even if an AI system is designed to advise human practitioners, rather than to carry out the actual diagnostic or treatment tasks, it may still result in detrimental unintended consequences, such as confirmatory bias and alert fatigue. A recent study showed that over-reliance on decision support systems resulted in increased false negative rate in radiology diagnoses, compared with the study scenario where the computer-aided diagnostic system was unavailable to the same group of radiologists.[14] Additionally, excessive warning information will result in alert fatigue,[15] and inexperienced practitioners may over-react to the warning messages. As such, AI developers need to pay attention to the clinical usage of automated systems, even if the systems only play an advisory role.

To address these challenges, researchers need to acknowledge and address the limitations in the current association-based machine learning paradigm and ensure quality control of the AI-based applications in various clinical settings and patient populations (table 1). For instance, the chest X-ray films with atypical feature statistics should be reviewed by radiologists to ensure that the obvious artefacts or unusual clinical contexts were adequately captured. In addition, prospective trials are needed to better understand the behaviour of AI systems in the real-world clinical settings, and continual calibration by human feedback is warranted to identify the development of emerging diseases as well as to examine the effectiveness of AI in recognising previously unclassified disease patterns. Due to the fact that AI technologies evolve at a fast pace and that machine learning models can update with additional pieces of information, regulatory bodies face a unique challenge in specifying trial requirements for regulatory approval.[16] To address this issue, the US Food and Drug Administration recently announced a pilot certification approach that inspects the AI developers, in addition to the product.[17] Detailed policies regarding the certification of developers are yet to be established. Furthermore, since many machine learning algorithms only focused on association identification, causal inference analyses are needed to characterise the causal relations underpinning the

BMJ Qual Saf: first published as 10.1136/bmjqs-2018-008551 on 5 October 2018. Downloaded from http://qualitysafety.bmj.com/ on 7 October 2018 by guest. Protected by copyright.

Table 1 A list of prominent issues of medical artificial intelligence (AI) applications and potential solutions

Issue	Potential solution
The Frame Problem (the difficulty in identifying and updating a set of axioms to properly describe the environment for autonomous agents).	▶ Clinician review of input with atypical feature statistics. ▶ Rigorous prospective clinical trials in diverse patient populations. ▶ Continual calibration by human feedback.
Trust in the performance of the AI program.	▶ Disclosure of implementation details, nature of training sets and shortcomings of the AI systems. ▶ Develop interpretable machine learning models. ▶ Patient education.
Amplifying biases presented in the historical data.	▶ Ongoing acquisition of training data reflecting current practice and population characteristics. ▶ Identify confounders in the association-based models.
Clinical workflow disruption.	▶ Redesign workflow that enables AI assistance without encouraging clinician decision-making passivity or aggravating 'alert fatigue'.

observed associations,[18] thereby mitigating the issues of confounding, and provide more transparency to the machine learning models. These steps are required to ensure public trust in novel medical AI applications.

Our focus on the near-term limitations of association-driven AI does not excuse any myopia regarding the highly variable and sometimes woefully inadequate performance of human clinicians. The mortality cost from medical errors alone, not including suboptimal decisions, has been widely documented for decades.[19] Nonetheless, since modern machine learning algorithms perform complex mathematical transformations to the input data,[16] errors made by computational systems will require extra vigilance to detect and interpret.[20] These cryptic errors and biases in the AI black boxes may systematically harm numerous patients simultaneously and worsen health disparities at scale.[21] In addition, even a robust AI application can reduce efficiency and cause additional medical errors if not adequately integrated into the current clinical workflow.[20] A better workflow would allow human clinicians and AI applications to compensate for their different and complementary weaknesses and blind spots to best serve the interests of patient safety and clinical efficiency.

Building an intelligent automated entity to evaluate, diagnose and treat patients in research settings is arguably the easiest part of designing an end-to-end medical AI system. In the context of the hype and hopes surrounding emerging AI applications in medicine, we need to acknowledge the brittleness of these systems, the importance of defining the correct frameworks for their application, and ensure rigorous quality control, including human supervision, to avoid driving our patients on autopilot towards unexpected, unwanted and unhealthful outcomes.

Contributors K-HY conceptualised and drafted the manuscript. ISK revised the manuscript and supervised the work.

Funding This study was funded by Harvard University (Harvard Data Science Fellowship) and the National Institutes of Health (Grant Number: OT3OD025466).

Competing interests Harvard Medical School submitted a provisional patent application on digital pathology profiling to the United States Patent and Trademark Office (USPTO).

Patient consent Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. Abrams R, Kurtz A, Brown J, 2016. Who died in self-driving accident, tested limits of his tesla: the New York Times. https://www.nytimes.com/2016/07/02/business/joshua-brown-technology-enthusiast-tested-the-limits-of-his-tesla.html (accessed 3 Aug 2018).
2. McCarthy J, Hayes PJ. Some philosophical problems from the standpoint of artificial intelligence. *Readings in Artificial Intelligence* 1969:431–50.
3. Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;316:2402–10.
4. Yu KH, Berry GJ, Rubin DL, *et al*. Association of omics features with histopathology patterns in lung adenocarcinoma. *Cell Syst* 2017;5:620–7.
5. Wang X, Peng Y, Lu L. ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *arXiv preprint arXiv* 2017:170502315.
6. Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115–8.
7. Sibbald B. Death but one unintended consequence of gene-therapy trial. *CMAJ* 2001;164:1612.
8. Wilson JM. Lessons learned from the gene therapy trial for ornithine transcarbamylase deficiency. *Mol Genet Metab* 2009;96:151–7.
9. Waltz E. After Theranos. *Nat Biotechnol* 2017;35:11–15.
10. Amin ANE, Parra MT, Kim-Farley R, *et al*. Ethical issues concerning vaccination requirements. *Public Health Rev* 2012;34:14.
11. Yu KH, Snyder M. Omics profiling in precision oncology. *Mol Cell Proteomics* 2016;15:2525–36.
12. Buranyi S, 2017. Rise of the racist robots – how AI is learning all our worst impulses the fuardian: the guardian. https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses (accessed 27 Mar 2018).
13. Richardson LD, Norris M. Access to health and health care: how race and ethnicity matter. *Mt Sinai J Med* 2010;77:166–77.
14. Lehman CD, Wellman RD, Buist DS, *et al*. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med* 2015;175:1828–37.
15. Phansalkar S, van der Sijs H, Tucker AD, *et al*. Drug-drug interactions that should be non-interruptive in order to reduce

BMJ Qual Saf: first published as 10.1136/bmjqs-2018-008551 on 5 October 2018. Downloaded from http://qualitysafety.bmj.com/ on 7 October 2018 by guest. Protected by copyright.

alert fatigue in electronic health records. *J Am Med Inform Assoc* 2013;20:489–93.
16. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng* 2018:(accepted).
17. U.S. Food and Drug Administration, 2018. Digital health software precertification (Pre-Cert) program. https://www.fda.gov/MedicalDevices/DigitalHealth/DigitalHealthPreCertProgram/default.htm (accessed 2 Aug 2018).
18. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health* 2006;60:578–86.
19. Stelfox HT, Palmisani S, Scurlock C, *et al*. The "To Err is Human" report and the patient safety literature. *Qual Saf Health Care* 2006;15:174–8.
20. Ash JS, Berg M, Coiera E. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. *J Am Med Inform Assoc* 2004;11:104–12.
21. O'Neil C. *Weapons of math destruction: how big data increases inequality and threatens democracy*. New York City: Broadway Books, 2016.

BMJ Qual Saf: first published as 10.1136/bmjqs-2018-008551 on 5 October 2018. Downloaded from http://qualitysafety.bmj.com/ on 7 October 2018 by guest. Protected by copyright.

VIEWPOINT

Artificial Intelligence in Health Care
Will the Value Match the Hype?

Ezekiel J. Emanuel, MD, PhD
Department of Medical Ethics and Health Policy, Perelman School of Medicine, University of Pennsylvania, Philadelphia; and Department of Health Care Management, Wharton School, University of Pennsylvania, Philadelphia.

Robert M. Wachter, MD
Department of Medicine, University of California, San Francisco.

Corresponding Author: Ezekiel J. Emanuel, MD, PhD, Perelman School of Medicine, University of Pennsylvania, 423 Guardian Dr, Blockley Hall, Ste 1412, Philadelphia, PA 19104 (mehpchair@upenn.edu).

Artificial intelligence (AI) and its many related applications (ie, big data, deep analytics, machine learning) have entered medicine's "magic bullet" phase. Desperate for a solution for the never-ending challenges of cost, quality, equity, and access, a steady stream of books, articles, and corporate pronouncements makes it seem like health care is on the cusp of an "AI revolution," one that will finally result in high-value care.

While AI has been responsible for some stunning advances, particularly in the area of visual pattern recognition,[1-3] a major challenge will be in converting AI-derived predictions or recommendations into effective action.

The most pressing problem with the US health care system is not a lack of data or analytics but changing the behavior of millions of patients and clinicians. Physician behaviors, including ordering tests, procedures, pharmaceuticals, and other treatments, are responsible for 80% of health care costs. Similarly, patient behaviors, including eating well, exercising, not smoking, moderate alcohol consumption, and medication adherence, influence more than half of the development of and outcomes related to chronic diseases. A narrow focus on data and analytics will distract the health system from what is needed to achieve health care transformation: meaningful behavior change.

Why the Hype?

It is unsurprising that AI would be the latest focus of health care hype. After all, AI, coupled with important changes in business models, underlies the disruption of industries ranging from retail to entertainment to finding transportation (eg, hailing a ride). Health care has been a laggard in these revolutions, largely because of the absence of a digital infrastructure. But that has changed. Although interoperability remains elusive and core digital tools, particularly electronic health records, are much maligned, the fact remains that health care is now collecting, storing, and moving data digitally. The so-called genetics revolution, and numerous precision medicine initiatives that are largely focused on storing and analyzing individual genetic information, have added to the massive amounts of data now available for analysis. In addition, there is interest in accessing the massive amounts of data from social media sites to help in the diagnosis and treatment of various disorders.

The business world and investor communities have noticed. Ten years ago, with much fanfare, Google and Microsoft both confidently ventured into health care only to experience sobering failure, which involved terminating (in the case of Google Health) and markedly scaling back (in the case of Microsoft HealthVault) their ambitious digital patient record initiatives. After these painful lessons, most digital giants did not get involved in health care. But that is no longer the situation. In the past few years, every major digital company has announced an AI-based health care initiative, with big dollar investments and the hiring of marquee talent.

Simultaneously, massive amounts of venture capital ($8.1 billion in 2018) are pouring into health care digital start-ups on the premise that health care is ripe for disruption, that AI is the tool to do it, and that the winning companies will reap untold profits.[4] It is a reasonable story, and early successes in fields ranging from insurance purchasing (Oscar, Bright) to management of chronic disease (Omada, Livongo) are fueling further investment.

The Crucial Need for an Effector Arm

The premise that more accurate and nuanced AI-based predictions will be transformative seems plausible, although that premise is likely wrong.

For example, the problem of translating evidence into practice has vexed the medical community since the evidence-based medicine movement began a generation ago. Why are only slightly more than half of evidence-based practices provided to patients?[5] Why does it take many years between the emergence of evidence that supports a new practice and consistent implementation of that practice?[6] Is it lack of accurate predictions? Ambiguity about the best course to take? No, neither of these are major factors.

To draw an analogy from immunology, the problem is ensuring that the effector arm functions efficiently and effectively. The body needs to identify foreign substances and organisms. But the crucial step is the activation of the immune system's effector arm—the antibody- and cell-mediated mechanisms, the complex array of cells, cytokines, complement, and more—that attack, neutralize, kill, and eliminate the intruders. Data, analytics, AI, and machine learning are about identification. But they have little role in establishing the structures, culture, and incentives necessary to change the behaviors of clinicians and patients.

For 30 years, physicians and others have tried various strategies to convert evidence about best practices into behavioral changes. For clinicians, the focus has been on education, practice guidelines, care paths, transparency, and incentives, mostly to little effect. Once the electronic health record era emerged, these strategies became digital and took the form of alerts, alarms, and order sets. However, new problems, such as alert fatigue and clinician frustration, have made clear how simplistic solutions are unlikely to be successful, even when delivered by expensive technologies rather than Post-it notes.

As for patients, consider the problem of low drug adherence. Only about 70% of all prescriptions are filled,

© 2019 American Medical Association. All rights reserved.

Downloaded From: https://jamanetwork.com/ by a Paris Sud University User on 05/20/2019

and of those that are filled only about 70% are taken properly for the full course of treatment.[7] Thus, half of individuals in the United States are nonadherent with medications, and the adherence rate is even lower for patients with chronic conditions with polypharmacy. Analyzing pharmacy data and other data sources to identify nonadherent patients or, better yet, using AI to predict which patients are likely to be nonadherent and relaying that information to their care team seems logical, but it is unlikely to reduce nonadherence substantially. Medicine needs to change how physicians and other clinicians interact with nonadherent patients and change patients' medication-taking habits. Simple tech approaches, like electronic pill caps, are unlikely to fix patient nonadherence.[8]

The gurus of data seem to assume that once something is identified and known, it is solved. That might be true in the tech world, where the aim is to hound consumers with electronic ads until they click on a link and buy a product. But in the health care system the goal is often changing an ingrained habit such as eating processed foods, smoking, not exercising, or skipping daily medications. There are no data to suggest that changing the precision of a prediction—such as, for example, explaining to a patient that "there's a very good chance your smoking will cause cancer or heart disease," compared with "there's a 27.6% chance your smoking will cause cancer or heart disease"—will succeed in changing behavior. The issue is the same when considering giving physicians more accurate predictions about the risk of readmission or sepsis. As Google indicated when it announced the closing of Google Health: "There has been adoption [of Google Health] among certain groups of users like tech-savvy patients.... But we haven't found a way to translate that limited usage into widespread adoption in *the daily health routines* of millions of people" (emphasis added).[9]

The Challenge of Behavior Change

Human beings are creatures of mental and physical habits. Changing those habits requires engagement and intentionality, and thus energy, sustained over months. This is why 80% of New Year's resolutions do not last past February.

There is a science to behavior change, and it is complex. It requires identifying triggers and changing the routine around them. It means not buying the packaged waffles but instead buying the yogurt and fruit. It could mean a patient taking medications and "rewarding" herself with morning coffee. All of this gets even more difficult when individuals are under stress.

In addition to changing patients' routines, physician and nurse routines also need to change, along with the routines and processes of care inside health care organizations. For example, consider how difficult it has been to change the simple routine of ensuring that clinicians thoroughly wash or sanitize their hands before examining patients.

A fundamental challenge facing the US health care system is to figure out how to effectively change routines and ensure these changes are embedded in the culture of the system. AI can have a role here, but it will not be simply through better predictions. Instead, the focus needs to be on the "effector arm of AI," thoughtfully combining the data with behavioral economics and other approaches to support positive behavioral changes. The change process will be iterative and messy, and there will be pushback. It will take place in hospitals and physician offices, not in Silicon Valley, although it is likely to require partnerships between tech companies and health care delivery organizations. Designing and implementing effector arms that induce meaningful behavior change will be the key to AI moving from the hype stage to one in which is it contributing to meaningful improvements in health and health care.

ARTICLE INFORMATION

Published Online: May 20, 2019. doi:10.1001/jama.2019.4914

Conflict of Interest Disclosures: Dr Emanuel reported receiving personal fees from Tanner Healthcare System, Mid-Atlantic Permanente Group, American College of Radiology, Marcus Evans, Loyola University, Oncology Society of New Jersey, Good Shepherd Community Care, Remedy Partners, Medzel, Kaiser Permanente Virtual Medicine, Wallace H. Coulter Foundation, Lake Nona Institute, Allocation, Philadelphia Chamber of Commerce, Blue Cross Blue Shield Minnesota, United Health Group, Futures Without Violence, Children's Hospital of Pennsylvania, Washington State Hospital Association, Association of Academic Health Centers, Blue Cross Blue Shield of Massachusetts, American Academy of Ophthalmology, Lumeris, Roivant Sciences Inc, Medical Specialties Distributors LLC, Vizient University Healthcare System, Center for Neuro-Degenerative Research, Colorado State University, Genentech Oncology Inc, Council of Insurance Agents and Brokers, Grifols Foundation, America's Health Insurance Plans, Montefiore Physician Leadership Academy, Greenwall Foundation, Medical Home Network, Healthcare Financial Management Association, Ecumenical Center–UT Health, American Association of Optometry, Associação Nacional de Hospitais Privados, National Alliance of Healthcare Purchaser Coalitions, Optum, Massachusetts Association of Health Plans, District of Columbia Hospital Association, and Washington University; holding stock in Gilead, Allergan, Amgen, Baxter, and United Health Group; and that he is a venture partner at Oak HC/FT. Dr Wachter reported serving on scientific advisory boards for PatientSafe Solutions, Early Sense, Amino, and Forward; being an investor in Smart Patients; serving on the board of directors for Accuity Medical Management; receiving personal fees from Commure, Teledoc, The Doctors Company, Nuance, GE Healthcare, Health Catalyst, AvaCare, and from approximately 50 nonprofit associations and healthcare organizations; and holding a patent to CareWeb with royalties paid.

REFERENCES

1. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402-2410. doi:10.1001/jama.2016.17216

2. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al; CAMELYON16 Consortium. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*. 2017;318(22):2199-2210. doi:10.1001/jama.2017.14585

3. Taylor AG, Mielke C, Mongan J. Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: a retrospective study. *PLoS Med*. 2018;15(11):e1002697. doi:10.1371/journal.pmed.1002697

4. Day S, Zweig M. 2018 Year-End Funding Report: is digital health in a bubble? Rock Health website. https://rockhealth.com/reports/2018-year-end-funding-report-is-digital-health-in-a-bubble/. 2019. Accessed May 9, 2019.

5. McGlynn EA, Asch SM, Adams J, et al. The quality of health care delivered to adults in the United States. *N Engl J Med*. 2003;348(26):2635-2645. doi:10.1056/NEJMsa022615

6. Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. *J R Soc Med*. 2011;104(12):510-520. doi:10.1258/jrsm.2011.110180

7. Brown MT, Bussell JK. Medication adherence: WHO cares? *Mayo Clin Proc*. 2011;86(4):304-314. doi:10.4065/mcp.2010.0575

8. Volpp KG, Troxel AB, Mehta SJ, et al. Effect of electronic reminders, financial incentives, and social support on outcomes after myocardial infarction: the HeartStrong randomized clinical trial. *JAMA Intern Med*. 2017;177(8):1093-1101. doi:10.1001/jamainternmed.2017.2449

9. An update on Google Health and Google PowerMeter. Google Blog website. https://googleblog.blogspot.com/2011/06/update-on-google-health-and-google.html. June 24, 2011. Accessed May 9, 2019.

© 2019 American Medical Association. All rights reserved.

Downloaded From: https://jamanetwork.com/ by a Paris Sud University User on 05/20/2019

VIEWPOINT

Addressing Bias in Artificial Intelligence in Health Care

Ravi B. Parikh, MD, MPP
Perelman School of Medicine, University of Pennsylvania, Philadelphia; and Corporal Michael J. Crescenz VA Medical Center, Philadelphia, Pennsylvania.

Stephanie Teeple, BA
Perelman School of Medicine, University of Pennsylvania, Philadelphia.

Amol S. Navathe, MD, PhD
Perelman School of Medicine, University of Pennsylvania, Philadelphia; and Corporal Michael J. Crescenz VA Medical Center, Philadelphia, Pennsylvania.

Author Audio Interview

Corresponding Author: Ravi B. Parikh, MD, MPP, Perelman School of Medicine, University of Pennsylvania, 423 Guardian Dr, Blockley 1102, Philadelphia, PA 19104 (ravi.parikh@pennmedicine.upenn.edu).

Recent scrutiny of artificial intelligence (AI)–based facial recognition software has renewed concerns about the unintended effects of AI on social bias and inequity. Academic and government officials have raised concerns over racial and gender bias in several AI-based technologies, including internet search engines and algorithms to predict risk of criminal behavior. Companies like IBM and Microsoft have made public commitments to "de-bias" their technologies, whereas Amazon mounted a public campaign criticizing such research. As AI applications gain traction in medicine, clinicians and health system leaders have raised similar concerns over automating and propagating existing biases.[1]

But is AI the problem? Or can it be part of the solution? While potentially inadvertently contributing to bias, AI technologies, when used responsibly, may also help counteract the risk of bias in unique ways. Using AI to identify bias in health care may help identify interventions that could help correct biased clinician decision-making and possibly reduce health disparities.

Statistical and Social Bias in AI

Statistical bias refers to an algorithm that produces a result that differs from the true underlying estimate. Statistical bias is common in predictive algorithms for many reasons, including suboptimal sampling, measurement error in predictor variables, and heterogeneity of effects. For example, the Framingham Study risk factors have been used for decades to predict risk of cardiovascular disease. However, the original Framingham Study sampled from an overwhelmingly non-Hispanic white population. When applying the Framingham Risk Score to populations with similar clinical characteristics, the predicted risk of a cardiovascular event was 20% lower for black individuals compared with white individuals, indicating that the score may not adequately capture risk factors for some minority groups.[2]

Social bias in health care refers to inequity in care delivery that systematically leads to suboptimal outcomes for a particular group. Social bias could be caused by a statistically biased algorithm or by other human factors, including implicit or explicit bias. For example, clinicians may incorrectly discount the diagnosis of myocardial infarction in older women because these patients are more likely to present with atypical symptoms.[3] An AI algorithm that learns from historical electronic health record (EHR) data and existing practice patterns may not recommend testing for cardiac ischemia for an older woman, delaying potentially lifesaving treatment. Perhaps of more concern, clinicians may be more likely to believe AI that reinforces current practice, thus perpetuating implicit social biases.

Why Do AI Algorithms Automate and Perpetuate Bias?

Artificial intelligence and machine learning are limited by the quality of data on which they are trained. The generalizability of AI algorithms across subgroups is critically dependent on factors like representativeness of included populations, missing data, and outliers. Generalizability and representativeness are also important considerations when interpreting randomized clinical trials.

However, the process by which the data are generated may be more important and particular to AI. If AI algorithms use data that are generated through a biased process, then the output may be similarly biased. This is a significant challenge when using clinical data sources like EHRs, insurance claims, or device readings because most of these data are generated as a consequence of human decisions. An algorithm to predict sepsis among patients in the emergency department, for example, may learn to use a test order for lactic acid to predict a poor outcome. However, the laboratory order may be more predictive of survival than the lactic acid value.[4] This is because a clinician is more likely to order the test for patients at risk of poor outcomes like death.

Artificial intelligence is also likely to incorrectly estimate risks for patients with missing data in the EHR. For example, among women with breast cancer, black women had a lower likelihood of being tested for high-risk germline mutations compared with white women, despite carrying a similar risk of such mutations.[5] Thus, an AI algorithm that depends on genetic test results is more likely to mischaracterize the risk of breast cancer for black patients than white patients.

While all predictive models may automate bias, AI may be unique in the extent to which bias is unrecognized (**Table**). Normally, clinicians have a pretest probability of an outcome and use the results of a diagnostic test to generate a posttest probability. However, clinicians may have a propensity to trust suggestions from AI decision support systems, which summarize large numbers of inputs into automated real-time predictions, while inadvertently discounting relevant information from nonautomated systems—so-called automation complacency.[6] For example, an AI-based early warning system can interpret changes in continuously monitored vital signs to alert an intensivist of a patient's impending clinical instability. However, AI-based decision support systems may produce a questionable or incorrect prediction. Hypothetically, an intensivist who is performing multiple concurrent tasks may inadvertently accept incorrect AI predictions unless there were obviously conflicting clinical information. This automation complacency could occur because AI predictions are framed around the outcome of interest and thus may be more salient to clinicians than an isolated test or laboratory result. Dedicated clinician training on interpreting AI outputs could ameliorate automation complacency.

Reducing Bias in AI

Although much of the discussion about AI and bias has focused on its potential for harm, strategies exist to mitigate such bias. When applied correctly, AI may be an effective tool to help counteract bias, an intractable problem in medicine.

© 2019 American Medical Association. All rights reserved.

Downloaded From: https://jamanetwork.com/ on 11/22/2019

Table. Artificial Intelligence Bias in Health Care

Example of Bias	Type of Bias	Potential Reasons for Bias	Methods to Address Bias
Low sensitivity of Framingham Risk Score in minority subgroups	Statistical	Algorithm training sample differs significantly from the population of interest	Oversample minority subgroups in training sample; tailor predictions or scores for specific subgroups
Delayed diagnosis of lung cancer in patients with low socioeconomic status or who lack transportation access to clinic	Social	Underlying disparities in diagnosis	Create flags for model uncertainty in predictions for certain high-risk subgroups
Missing data in electronic health record-based data sets due to lack of patient follow-up	Statistical and social	Missing data	Base predictions on "upstream" data at presentation of illness, not on subsequent follow-up data

First, AI decision support tools could be used to identify real-time bias in physician decision-making. Many nonmedical factors affect physician decision-making; situations with high cognitive load, such as decision-making at the end of a clinic day, are particularly prone to bias. If rational AI predictions and clinician decision-making differ in these situations, clinicians could be alerted in real time about decisions that are at risk of bias. For example, an AI algorithm may flag a possibly questionable opioid prescription at the end of a primary care clinician's day, providing a needed check on this decision. There are fledgling examples of using AI to identify disparities. When applied to unstructured data from psychiatry notes, AI algorithms demonstrated greater documentation of anxiety and chronic pain topics for white patients and psychosis topics for black, Hispanic, and Asian patients. Alerting clinicians to these disparities in documentation in real time could improve care of patients by making implicit biases in their practice more salient.[7]

Second, because most AI bias is related to the data-generating process, the primary solution may be to preferentially use unbiased data sources. Uniform collection of large amounts of data on all patients is now possible because of more routine use of noninvasive monitoring. Examples of relatively unbiased, uniform data sources include recorded vital sign data during surgical operations or triage data collected from the first hour after emergency department presentation, "upstream" of clinician judgments. Randomized trial data also could be used preferentially instead of observational data to support AI development, although it would be important to access which patients had been enrolled in the clinical trials. In many regards, the potential bias in AI is similar to concerns raised in clinical trials, in that participants are often nonrepresentative of the general population.

Other steps could help facilitate addressing bias in health care AI. For instance, existing standards, including the PROBAST tool to assess risk of bias in prediction models, can aid algorithm developers in selecting representative training sets and appropriate predictor variables.[8] In addition, algorithm predictions and subsequent actions could be tracked continuously to help ensure that outputs are not reinforcing existing social biases. Algorithm developers also could use certain sensitivity checks, including creating simulated data sets with high numbers of omitted variables and conducting counterfactual simulations, to determine how robust predictions are to omitted variable bias. For data sets that are necessarily collected after clinician decisions, algorithm developers could seek to oversample underrepresented populations to mitigate statistical bias.

Conclusions

Artificial intelligence is making its way into clinical practice. Because of its reliance on historical data, which are based on biased data generation or clinical practices, AI can create or perpetuate biases that may worsen patient outcomes. However, by strategically deploying AI and carefully selecting underlying data, algorithm developers can mitigate AI bias. Addressing bias could allow AI to reach its fullest potential by helping to improve diagnosis and prediction while protecting patients.

ARTICLE INFORMATION

Published Online: November 22, 2019. doi:10.1001/jama.2019.18058

Conflict of Interest Disclosures: Dr Parikh reported receipt of personal fees from GNS Healthcare. Dr Navathe reported receipt of grants from the Hawaii Medical Service Association, the Anthem Public Policy Institute, the Commonwealth Fund, Oscar Health, Cigna Corporation, and the Donaghue Foundation and personal fees from Navvis Healthcare, Agathos Inc, University Health System (Singapore), Elsevier Press, Navahealth, the Cleveland Clinic, the Medicare Payment Advisory Commission, and Embedded Healthcare; he reported being an uncompensated board member for Integrated Services Inc. No other disclosures were reported.

Funding/Support: This work was supported in part by the Penn Center for Precision Medicine (Dr Parikh) and the Pennsylvania Universal Research Enhancement (CURE) Program and Robert Wood Johnson Foundation (Dr Navathe).

Role of the Funders/Sponsors: The funders had no role in the preparation, review, or approval of the manuscript or decision to submit the manuscript for publication.

REFERENCES

1. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med*. 2018;178(11):1544-1547. doi:10.1001/jamainternmed.2018.3763

2. Gijsberts CM, Groenewegen KA, Hoefer IE, et al. Race/ethnic differences in the associations of the Framingham risk factors with carotid IMT and cardiovascular events. *PLoS One*. 2015;10(7):e0132321. doi:10.1371/journal.pone.0132321

3. Canto JG, Goldberg RJ, Hand MM, et al. Symptom presentation of women with acute coronary syndromes: myth vs reality. *Arch Intern Med*. 2007;167(22):2405-2413. doi:10.1001/archinte.167.22.2405

4. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ*. 2018;361:k1479. doi:10.1136/bmj.k1479

5. McCarthy AM, Bristol M, Domchek SM, et al. Health care segregation, physician recommendation, and racial disparities in *BRCA1/2* testing among women with breast cancer. *J Clin Oncol*. 2016;34(22):2610-2618. doi:10.1200/JCO.2015.66.0019

6. Parasuraman R, Manzey DH. Complacency and bias in human use of automation: an attentional integration. *Hum Factors*. 2010;52(3):381-410. doi:10.1177/0018720810376055

7. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? *AMA J Ethics*. 2019;21(2):E167-E179. doi:10.1001/amajethics.2019.167

8. Wolff RF, Moons KGM, Riley RD, et al; PROBAST Group. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51-58. doi:10.7326/M18-1376

© 2019 American Medical Association. All rights reserved.

Downloaded From: https://jamanetwork.com/ on 11/22/2019

VIEWPOINT

Artificial Intelligence in Health Care
A Report From the National Academy of Medicine

Michael E. Matheny, MD, MS, MPH
Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee; and Geriatric Research Education & Clinical Care Service, Tennessee Valley Healthcare System VA, Nashville.

Danielle Whicher, PhD, MHS
Mathematica Policy Research, Washington, DC.

Sonoo Thadaney Israni, MBA
Stanford University School of Medicine, Stanford, California.

The promise of artificial intelligence (AI) in health care offers substantial opportunities to improve patient and clinical team outcomes, reduce costs, and influence population health. Current data generation greatly exceeds human cognitive capacity to effectively manage information, and AI is likely to have an important and complementary role to human cognition to support delivery of personalized health care.[1] For example, recent innovations in AI have shown high levels of accuracy in imaging and signal detection tasks and are considered among the most mature tools in this domain.[2]

However, there are challenges in realizing the potential for AI in health care. Disconnects between reality and expectations have led to prior precipitous declines in use of the technology, termed *AI winters*, and another such event is possible, especially in health care.[3] Today, AI has outsized market expectations and technology sector investments. Current challenges include using biased data for AI model development, applying AI outside of populations represented in the training and validation data sets, disregarding the effects of possible unintended consequences on care or the patient-clinician relationship, and limited data about actual effects on patient outcomes and cost of care.

AI in Healthcare: The Hope, The Hype, The Promise, The Peril, a publication by the National Academy of Medicine (NAM), synthesizes current knowledge and offers a reference document for the responsible development, implementation, and maintenance of AI in the clinical enterprise.[4] The publication outlines current and near-term AI solutions; highlights the challenges, limitations, and best practices for AI development, adoption, and maintenance; presents an overview of the legal and regulatory landscape for health care AI; urges the prioritization of equity, inclusion, and a human rights lens for this work; and outlines considerations for moving forward. This Viewpoint shares highlights from the NAM publication.

> Health care is at a critical juncture for the safe and effective use of AI algorithms and tools in supporting the health of patients.

Promoting population-representative data with accessibility, standardization, and quality is imperative. Health care AI should be trained and validated on population-representative data to ensure accuracy for all populations and to achieve performance levels necessary for scalable success. Trends such as decreasing cost for storing and managing data, data collection via electronic health records, and exponential consumer health data generation, have created a data-rich health care ecosystem. However, there continue to be issues of data quality, appropriate consent, interoperability, and scale of data transfers. The current challenges are grounded in patient and health care system preferences, regulations, and political will rather than technical capacity or specifications. It is prudent to engage AI developers, users, and patients and their families in discussions about appropriate policy, regulatory, and legislative solutions.

Prioritize ethical, equitable, and inclusive health care AI while addressing explicit and implicit bias. This should be a clearly stated goal when developing and deploying tools in consumer and clinical settings. Today's health care inequities include societal bias, social determinants of health, and perverse incentives in the existing system. Further exacerbating the lack of trust are high-profile, biased AI deployed for judicial sentencing, facial recognition, and hiring practices.[5] It is essential to ascertain the applicability of the data used to develop AI by scrutinizing the underlying biases to understand its potential to worsen or address existing inequities, and whether and how it should be deployed.[6] Leveraging diverse data sets is essential, as is preventing unintended consequences resulting from privacy breaches and inappropriate deployment. A quintuple aim should be the goal, adding equity and inclusion to the quadruple aim of improving the health of the population, enhancing the patient experience, reducing per capita cost, and enhancing clinician wellness.

Contextualizing the dialogue of transparency and trust requires accepting differential needs. Full transparency with respect to the population-representativeness, composition, semantics, provenance, and quality of data used to develop AI tools is critical. There also needs to be full transparency and assessment of relevant performance components of AI. However, algorithmic transparency should not be required for all use cases. AI developers, implementers, users, and regulators should collaboratively define guidelines for clarifying the level of transparency needed across a spectrum. There should be a clear separation of data, performance, and algorithmic transparency.

Near-term focus is needed on augmented intelligence vs AI autonomous agents. Fully autonomous AI is inciting public concern and faces numerous technical and regulatory challenges. Realistically, the current opportunity is *augmented intelligence*, supporting data synthesis, interpretation, and decision-making for

Corresponding Author: Michael E. Matheny, MD, MS, MPH, Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Ave, Nashville, TN 37212 (michael.matheny@vanderbilt.edu).

© 2019 American Medical Association. All rights reserved.

Downloaded From: https://jamanetwork.com/ on 12/17/2019

clinicians, allied health professionals, and patients. Focusing on this reality is essential for developing user trust because there is an understandable low tolerance for machine error, and these tools are being implemented in an environment of inadequate regulation and legislation.

Develop and deploy appropriate training and educational programs to support health care AI. The scale at which AI may change the landscape of prevention, diagnosis, treatment, and health care management is substantial. The curricula must be multidisciplinary and engage AI developers, implementers, health care system leadership, frontline clinical teams, ethicists, humanists, patients, and caregivers. Each group brings much needed perspectives, requirements, and expertise. Data science curricula should expand to include teaching how engaging diverse development teams is likely to improve the utility and effect of AI, and also to raise the awareness of ethics, equity, inclusion, and potential unintended consequences. Health care professional training should incorporate curricula on how to appropriately assess and use AI products and services. Adding these components via continuing education for current practitioners in all relevant fields should be a priority. Consumer health educational programs, at all educational levels, are needed to help inform consumers about consent, privacy, and health care AI savviness.

Leverage frameworks and best practices for learning health care systems, human factors, and implementation science to address the challenges in operationalizing health care AI. The AI community should develop guidance on best practices for inclusivity and equity, software development, implementation science, and human-computer interaction, all within the framework of the learning health care system. Health care delivery systems should have a robust and mature information technology (IT) governance strategy prior to embarking on substantial AI deployment and integration. In addition, a national focus on providing appropriate health care AI in resource constrained environments is needed.

Balance innovation with safety via regulation and legislation to promote trust. AI has the potential to improve patient outcomes but could also pose significant risks in terms of inappropriate or inaccurate patient risk assessment, treatment recommendations, diagnostic error, privacy breaches, and other factors. While regulators should remain flexible, the potential for lagging legal responses will remain a challenge for AI innovation. Recent congressional and US Food and Drug Administration developments and guidance have made progress, and it is important to pursue a graduated approach based on levels of patient risk and AI autonomy, including considerations for static or dynamic AI. Liability will continue to evolve as regulators, courts, and the risk-management industries weigh in, and a careful balance and understanding of this is critical for AI adoption.[7] Regulators and patients and their families should encourage AI developers, health system leaders, clinical users, and informatics and health IT experts to evaluate deployed clinical AI for effectiveness and safety based on clinical data.

Conclusions

Health care is at a critical juncture for the safe and effective use of AI algorithms and tools in supporting the health of patients. The technical capacity exists to leverage these tools to transform health care. The challenges are unrealistic expectations, biased and nonrepresentative data, inadequate prioritization of equity and inclusion, the risk of exacerbating health care disparities, low levels of trust, uncertain regulatory and tort environments, and inadequate evaluation before scaling narrow AI.

ARTICLE INFORMATION

Published Online: December 17, 2019. doi:10.1001/jama.2019.21579

Conflict of Interest Disclosures: Ms Thadaney Israni reports being supported by Presence (a Stanford Medicine Center, Stanford University). Dr Matheny reports receiving grants from the Veterans Administration. No other disclosures were reported.

Disclaimer: This Viewpoint provides a summary of a publication developed as part of the National Academy's Digital Learning Collaborative Initiative within the Leadership Consortium: Collaboration for a Value & Science-Driven Learning Health System (https://nam.edu/programs/value-science-driven-health-care/digital-learning/). This article reflects the views of leading authorities on the issues engaged and does not represent formal consensus positions of the National Academy of Medicine or the organizations of the participating authors.

Additional Contributions: Coauthors of the National Academy of Medicine publication include the following: Michael McGinnis, MD, MA, MPP, Jonathan B. Perlin, MD, PhD, Reed Tuckson, MD, Mahnoor Ahmed, MEng, Paul Bleicher, MD, PhD, Wendy Chapman, PhD, Jim Fackler, MD, Edmund Jackson, PhD, Joachim Roski, PhD, MPH, Jaimee Heffner, PhD, Ranak Trivedi, PhD, Guilherme Del Fiol, MD, PhD, Rita Kukafka, DrPh, Hossein Estiri, PhD, Joni Pierce, MBA, Jeffrey Klann, PhD, Jonathan Chen, MD, PhD, Andrew Beam, PhD, Suchi Saria, PhD, Eneida A. Mendonca, MD, PhD, Hongfang Liu, PhD, Jenna Wiens, PhD, Anna Goldberg, PhD, Nigam Shah, MBBS, PhD, Stephan Fihn, MD, MPH, Seth Hain, MS, Andrew Auerbach, MD, Douglas McNair, MD, PhD, and Nicholson Price, JD, PhD; they received no compensation.

REFERENCES

1. National Research Council. *Computational Technology for Effective Health Care: Immediate Steps and Strategic Directions*. Washington, DC: National Academies Press; 2009.

2. Liu F, Zhou Z, Samsonov A, et al. Deep learning approach for evaluating knee MR Images: achieving high diagnostic performance for cartilage lesion detection. *Radiology*. 2018;289(1):160-169. doi:10.1148/radiol.2018172986

3. Newquist H. *The Brain Makers: Genius, Ego, and Greed in the Search for Machines That Think*. Indianapolis, IN: Sams Publishing; 1994.

4. Matheny ME, Thadaney Israni S, Ahmed M, Whicher D. *AI in Health Care: The Hope, the Hype, the Promise, the Peril*. Washington, DC: National Academy of Medicine; 2019. https://nam.edu/artificial-intelligence-special-publication.

5. Dastin J. Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G. October 9, 2018. Accessed December 13, 2019.

6. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care [published online November 23, 2019]. *JAMA*. doi:10.1001/jama.2019.18058

7. Price WN II, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. *JAMA*. 2019;322(18):1765-1766. doi:10.1001/jama.2019.15064

© 2019 American Medical Association. All rights reserved.

Downloaded From: https://jamanetwork.com/ on 12/17/2019

Artificial intelligence in medicine:

current trends and future possibilities

Artificial intelligence (AI) research within medicine is growing rapidly. In 2016, healthcare AI projects attracted more investment than AI projects within any other sector of the global economy.[1] However, among the excitement, there is equal scepticism, with some urging caution at inflated expectations.[2] This article takes a close look at current trends in medical AI and the future possibilities for general practice.

WHAT IS MEDICAL ARTIFICIAL INTELLIGENCE?

Informing clinical decision making through insights from past data is the essence of evidence-based medicine. Traditionally, statistical methods have approached this task by characterising patterns within data as mathematical equations, for example, linear regression suggests a 'line of best fit'. Through 'machine learning' (ML), AI provides techniques that uncover complex associations which cannot easily be reduced to an equation. For example, neural networks represent data through vast numbers of interconnected neurones in a similar fashion to the human brain. This allows ML systems to approach complex problem solving just as a clinician might — by carefully weighing evidence to reach reasoned conclusions. However, unlike a single clinician, these systems can simultaneously observe and rapidly process an almost limitless number of inputs. For example, an AI-driven smartphone app now capably handles the task of triaging 1.2 million people in North London to Accident & Emergency (A&E).[3] Furthermore, these systems are able to learn from each incremental case and can be exposed, within minutes, to more cases than a clinician could see in many lifetimes. This is why an AI-driven application is able to out-perform dermatologists at correctly classifying suspicious skin lesions[4] or why AI is being trusted with tasks where experts often disagree, such as identifying pulmonary tuberculosis on chest radiographs.[5] Although AI is a broad field, this article focuses exclusively on ML techniques because of their ubiquitous usage in important clinical applications.

WHAT ARE THE CURRENT TRENDS IN MEDICAL AI?

Aside from simply demonstrating superior efficacy, new technologies entering the medical field must also integrate with current practices, gain appropriate regulatory approval, and, perhaps most importantly, inspire medical staff and patients to invest in a new paradigm. These challenges have led to a number of emerging trends in AI research and adoption.

> *"... these systems are able to learn from each incremental case and can be exposed, within minutes, to more cases than a clinician could see in many lifetimes."*

AI excels at well-defined tasks

Research has focused on tasks where AI is able to effectively demonstrate its performance in relation to a human doctor. Generally, these tasks have clearly defined inputs and a binary output that is easily validated. In classifying suspicious skin lesions, the input is a digital photograph and the output is a simple binary classification: benign or malignant. Under these conditions, researchers simply had to demonstrate that AI had superior sensitivity and specificity than dermatologists when classifying previously unseen photographs of biopsy-validated lesions.[4]

AI is supporting doctors, not replacing them

Machines lack human qualities such as empathy and compassion, and therefore patients must perceive that consultations are being led by human doctors. Furthermore, patients cannot be expected to immediately trust AI; a technology shrouded by mistrust.[6] Therefore, AI commonly handles tasks that are essential, but limited enough in their scope so as to leave the primary responsibility of patient management with a human doctor. There is an ongoing clinical trial using AI to calculate target zones for head and neck radiotherapy more accurately and far more quickly than a human being. An interventional radiologist is still ultimately responsible for delivering the therapy but AI has a significant background role in protecting the patient from harmful radiation.[7]

AI supports poorly resourced services

A single AI system is able to support a large population and therefore it is ideally suited to situations where human expertise is a scarce resource. In many TB-prevalent countries there is a lack of radiological expertise at remote centres.[8] Using AI, radiographs uploaded from these centres could be interpreted by a single central system; a recent study shows that AI correctly diagnoses pulmonary TB with a sensitivity of 95% and specificity of 100%.[5] Furthermore, under-resourced tasks where patients are experiencing unsatisfactory waiting times are also attractive to AI in the form of triage systems.[3]

AI is a very picky eater

Developing ML models requires well-structured training data about a phenomenon that remains relatively stable over time. A departure from this results in 'over-fitting', where AI gives undue importance to spurious correlations within past data. In 2008, Google tried to predict the seasonal prevalence of influenza using only

> *"... AI commonly handles tasks that are essential, but limited enough in their scope so as to leave the primary responsibility of patient management with a human doctor."*

"Integrating these systems into clinical practice necessitates building a mutually beneficial relationship between AI and clinicians ..."

the search terms entered into its search engine. Because people's searching habits change dramatically with every passing year, the model was so poorly predictive of the future that it was quickly discontinued.[9] Additionally, data that are anonymised and digitised at source are also preferable, as this aids in research and development.

FUTURE POSSIBILITIES IN GENERAL PRACTICE

AI will extract important information from a patient's electronic footprint. At first this will save time and improve efficiency, but following adequate testing it will also directly guide patient management. Take the example of a consultation with a patient with type 2 diabetes; currently a clinician spends significant time reading outpatient letters, checking blood tests, and finding clinical guidelines from a number of disconnected systems. In contrast, AI could automatically prepare the most important risks and actions given the patient's clinical record. It could also automatically convert recorded dialogue of the consultation into a summary letter for the clinician to approve or amend. Both of these applications would save considerable time and could be implemented very quickly because they assist clinicians rather than replacing them.

As these systems become better validated, they will be given more responsibility. For the patient with type 2 diabetes, the threshold of statin commencement could be determined by AI on an individualised basis given nuisances of the patient's history rather than a rigidly defined 'one-size-fits-all' algorithm. The research required for this 'personalised' medicine would only be possible through AI intelligently summarising enormous quantities of medical information. Furthermore, because AI is able to simultaneously monitor millions of inputs, it will have a significant role in preventative medicine. AI could proactively suggest consultations when it determines that the patient's risk of developing a particular diabetic complication warrants intervention. In contrast, it would be impractical to task a human being with the responsibility of closely monitoring every test result and appointment of every diabetic patient in a practice in real time.

AI-based systems will also bring specialist diagnostic expertise into primary care. If an image of a skin lesion is sufficient to capably diagnose its aetiology, images could be captured at a GP practice and sent to a specialist dermatology AI system for instant analysis. Patients identified as low risk would receive instant reassurance while high-risk patients would experience lower referral waiting times because clinics would only be receiving selected cases. This concept is not limited to skin lesions, AI has shown potential in interpreting many different types of image data including retinal scans,[10] radiographs,[5] and ultrasound.[11] Many of these images can be captured with relatively inexpensive and widely available equipment.

Future AI research should be directed towards carefully selected tasks that broadly align with the trends outlined in this article. Integrating these systems into clinical practice necessitates building a mutually beneficial relationship between AI and clinicians, where AI offers clinicians greater efficiency or cost-effectiveness and clinicians offer AI the essential clinical exposure it needs to learn complex clinical case management. Throughout the process it will be critical to ensure that AI does not obscure the human face of medicine because the biggest impediment to AI's widespread adoption will be the public's hesitation to embrace an increasingly controversial technology.[12]

Varun H Buch,
Clinical Technology Lead, Cera Care, London.

Irfan Ahmed,
GP Lead, Cera Care, London.

Mahiben Maruthappu,
CEO, Cera Care, London.

Provenance
Freely submitted; externally peer reviewed.

Competing interests
The authors have declared no competing interests.

DOI: https://doi.org/10.3399/bjgp18X695213

ADDRESS FOR CORRESPONDENCE

Mahiben Maruthappu
Cera Care, 219 Kensington High Street, Kensington, London W8 6BD, UK.

E-mail: mahiben.maruthappu@nhs.net

REFERENCES

1. CB Insights Research. Healthcare remains the hottest AI category for deals. 2017. https://www.cbinsights.com/research/artificial-intelligence-healthcare-startups-investors/ (accessed 15 Jan 2018).
2. Chen JH, Asch SM. Machine learning and prediction in medicine — beyond the peak of inflated expectations. *N Eng J Med* 2017; **376(26):** 2507–2509.
3. Burgess M. The NHS is trialling an AI chatbot to answer your medical questions. *Wired* 2017: **5 Jan:** http://www.wired.co.uk/article/babylon-nhs-chatbot-app (accessed 15 Jan 2018).
4. Esteva A, Kuprel B, Novoa RA, *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542(7639):** 115–118.
5. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* 2017; **284(2):** 574–582.
6. Oppenheim M. Stephen Hawking: artificial intelligence could be the greatest disaster in human history. *Independent* 2016; **20 Oct:** http://www.independent.co.uk/news/people/stephen-hawking-artificial-intelligence-diaster-human-history-leverhulme-centre-cambridge-a7371106.html (accessed 15 Jan 2017).
7. Chu C, De Fauw J, Tomasev N, *et al.* Applying machine learning to automated segmentation of head and neck tumour volumes and organs at risk on radiotherapy planning CT and MRI scans. *F1000Research* 2016; **5:** 2104.
8. Hoog AH, Meme HK, van Deutekom H, *et al.* High sensitivity of chest radiograph reading by clinical officers in a tuberculosis prevalence survey. *Int J Tuberc Lung Dis* 2011; **15(10):** 1308–1314.
9. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google flu: traps in big data analysis. *Science* 2014; **343(6176):** 1203–1205.
10. Shead S. Google DeepMind is funding NHS research at Moorfields Eye Hospital. *Business Insider* 2017; **3 Aug:** http://uk.businessinsider.com/deepmind-is-funding-nhs-research-2017-7 (accessed 15 Jan 2018).
11. Chen H, Wu L, Dou Q, *et al.* Ultrasound standard plane detection using a composite neural network framework. *IEEE Trans Cybern* 2017; **47(6):** 1576–1586.
12. Naughton J. Giving Google our private NHS data is simply illegal. *Guardian* 2017; **9 Jul:** https://www.theguardian.com/commentisfree/2017/jul/09/giving-google-private-nhs-data-is-simply-illegal (accessed 15 Jan 2018).

VIEWPOINT

Humanizing Artificial Intelligence

Sonoo Thadaney Israni, MBA
Stanford University, Stanford, California.

Abraham Verghese, MD
Stanford University, Stanford, California.

Viewpoint

Author Audio Interview

If human intelligence is the learned ability to gain from experience and the capacity to handle unfamiliar situations and manipulate abstract concepts while using experience and knowledge to change the world, then the concept of artificial intelligence (AI)—a huge advance in data processing and computing—would not easily compare with true human intelligence. In this Viewpoint, AI broadly encompasses machine learning, natural language processing, expert systems that emulate the decision making and reasoning of human experts, and other related applications.

The promise of AI is undeniable; it is possible that the hype and fear surrounding the subject are greater than that which accompanied the discovery of the structure of DNA or the whole genome. AI tools that recognize images more accurately and consistently than humans can are exciting advances for clinicians. But, the recurring trope of pitting humans vs machine (eg, Deep Blue vs Gary Kasparov) misses the point. Gilder anticipated the effects of many new technologies, noting that "Machines can't be minds.... Creativity always comes as a surprise to us. If it wasn't surprising, we wouldn't need it. Machines are not capable of creativity. Human minds can generate counterfactuals, imaginative flights, dreams. By contrast, a surprise in a machine is a breakdown."[1] Machines do create and surprise in art and music, but taste for this sort of originality is uncertain. Surprises in the fine arts create reflection; surprises in medical diagnosis or treatment are unwelcome.

> Flawed or incomplete data sets that are not inclusive can automate inequality.

In discussing the prospects of AI within health care, 1 opportunity is often overlooked: could AI help clinicians deliver better and more humanistic care? Beyond easing the cognitive load and, at times, the drudgery of a busy practice, can AI help clinicians become better at being human? The desirable attributes of humans who choose the path of caring for others include, in addition to scientific knowledge, the capacity to love, to have empathy, to care and express caring, to be generous, to be brave in advocating for others, to do no harm, and to work for the greater good and advocate for justice. How might AI help clinicians nurture and protect these qualities? This type of challenge is rarely discussed or considered at conferences on AI and medicine, perhaps because it is viewed as messy and hard to define. But, if the goal is for AI to emulate the best qualities of human intelligence, it is precisely the territory that cannot be avoided.

Forethought in the application of new technology to prevent and predict unintended consequences that become apparent over time is a much-needed and exciting new field of inquiry. Such forethought might have anticipated the angst, depression, and disenfranchisement physicians felt when the Health Information Technology for Economic and Clinical Health Act and its mandates resulted in 2 leading electronic health record systems (EHRs) emerging as dominant tools, both designed for maximizing revenue in complex hospital environments. These systems were not designed to help care for patients or to ease the burden of those who do, even though they are a great advance on what preceded the EHR. The billing component works well, but the clinical notes are an admixture of cut-and-paste notes and templates that stretch the truth and erode practitioner morale. Imperfect solutions and workarounds that have evolved include scribes, but it goes against the grain of progress to reach for a solution that was fitting for the Roman Empire. Creative innovators are surely capable of something better. Scribes—modern or ancient—can be seen as a metaphor for what is really needed: a nonintrusive entity that transforms clinicians from the highest-paid scribe to someone who is unimpeded in using his or her skills and years of training to care for the patient.

The great variability of human beings is what makes medicine an art. As Osler observed, it is "more important to know what sort of a patient has a disease than what sort of a disease a patient has."[2] If AI can help with a more astute knowledge of the patient and the "framily" (ie, unpaid caregivers, who are friends and family), it would be the kind of advance that could help clinicians become better at delivering more humanistic care. For instance, if an AI-powered EHR could prepare the clinical team with previsit material beyond the rote medical and family or environmental and social history, digested in a vivid useable form with graphics and animation equivalent to what is readily available in other spheres of the digital world, it would be possible for physicians to picture precisely where this patient is in his or her life. If AI, natural language processing, and video captured what actually transpired during the clinical encounter, could the clinician prioritize the patient and family during the visit? Such advances might allow clinicians more time to engage face-to-face (and not just electronically) with colleagues in the shared enterprise of caring for a unique patient.

Crucially, if AI is going to make clinicians better at caring for humans in distress, the data sets being used must be representative of society and not biased by sex, race, ethnicity, socioeconomic status, age, ability, and geography.[3] This need for representation is not only a data science issue, but also a moral one. In the absence of equal representation, society has already seen inequitable criminal justice sentencing, unfair hiring practices, and loan-risk determination, to name a few injustices.[4] A 2018 revised study of pooled cohort

Corresponding Author: Abraham Verghese, MD, Department of Medicine, Stanford University, 300 Pasteur Dr, S102, Stanford, CA 94305-5110 (Abrahamv@stanford.edu).

© 2018 American Medical Association. All rights reserved.

Downloaded From: by a Göteborgs Universitet User on 12/10/2018

equations[5] found that African American individuals were underrepresented in the initial sample of people that was used to create the 2013 guidelines for cardiovascular risk, thereby overestimating risk of cardiovascular disease compared in that population. Flawed or incomplete data sets that are not inclusive can automate inequality.[6]

On the human side, data scientists, social scientists, computer scientists, and clinicians must also reflect in their composition the society they hope to serve; their life experiences are critical in authoring and building meaningful products. For instance, in 2014, the Healthkit was billed by Apple as a tool to help track blood alcohol content, height, inhaler use, sodium intake, and other parameters, so that "you can monitor all of your metrics that you're most interested in."[7] But, the tool did not track a woman's menstrual cycle, and many critics pointed to the lack of female engineers as a possible cause of this embarrassing oversight.[7]

Humans in need want the best of what science and medicine have to offer. Particularly, in the setting of serious and chronic illnesses, patients in need want their physicians to be human beings who care, communicate clearly, and are compassionate and express empathy. As Peabody famously observed long ago, "One of the essential qualities of the clinician is interest in humanity, for the secret of the care of the patient is in caring for the patient."[8] Bettering the ability of physicians to truly care for and express caring is the challenge for colleagues in computer science and medical informatics. Systems that augment the diagnostic and scientific task of treating disease are exciting and wonderful, but is it possible to invent and discover applications that can enhance the human abilities in clinicians to better engage in caring for the patient? This possibility would be a significant breakthrough. Many people are hopeful that it is just the sort of breakthrough that intelligent humans can achieve.

ARTICLE INFORMATION

Published Online: December 10, 2018. doi:10.1001/jama.2018.19398

Conflict of Interest Disclosures: Ms Thadaney has no conflicts to report. Dr Verghese receives royalties from Simon & Schuster and Random House publishers, serves on the Gilead Health Policy Advisory Board, and the Leigh Speakers Bureau.

Additional Contributions: The authors are grateful to Stanford University's Jonathan H. Chen, MD, PhD; Roberta R. Katz, JD, PhD; and Nigam H. Shah, MBBS, PhD, for critically reviewing this paper.

REFERENCES

1. Varadarajan T. Sage against the machine. Wall Street Journal. https://www.wsj.com/articles/sage-against-the-machine-1535747443. August 31, 2018. Accessed October 23, 2018.

2. Silverman M, Murray TJ, Bryan CS, eds. *The Quotable Osler*. Philadelphia, PA: American College of Physicians; 2007.

3. Caplan A, Friesen P. Health disparities and clinical trial recruitment: is there a duty to tweet? PLOS Biology. 2017;15(3):e2002040. doi:10.1371/journal.pbio.2002040

4. ONeil C. *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*. New York, NY: Penguin Books; 2016.

5. Yadlowsky S, Hayward RA, Sussman JB, McClelland RL, Min Y-I, Basu S. Clinical implications of revised pooled cohort equations for estimating atherosclerotic cardiovascular disease risk. *Ann Intern Med*. 2018;169(1):20-29. doi:10.7326/M17-3011

6. Virginia E. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* New York, NY: St Martin's Press; 2018.

7. Eveleth R. How self-tracking apps exclude women. *The Atlantic*. https://www.theatlantic.com/technology/archive/2014/12/how-self-tracking-apps-exclude-women/383673/ December 15, 2014. Accessed December 5, 2018.

8. Peabody F. The care of the patient. *JAMA*. 1927;88(12):877-882. doi:10.1001/jama.1927.02680380001001

© 2018 American Medical Association. All rights reserved.

Downloaded From: by a Göteborgs Universitet User on 12/10/2018

منابع
آزمون مرحله دوم
انفرادی
و
آزمونهای گروهی

❊❊❊❊

این منابع در آزمون غربالگری (مرحله اول انفرادی) مورد استفاده قرار نمیگیرند و از آنها سوالی طرح نخواهد شد

Artificial Intelligence in Medicine?

21.0 Introduction

Among all types of machines used in medicine today, the computing machinery plays the predominant role. This also includes its indispensable contribution to the operation of nearly all modern devices used in medical practice and research, e.g., X-ray machines, MRT, heart-lung machines, dialysis machines, surgical robots, pacemakers, and others. There is almost no medical device today without specific chips and computing components. So, the question arises how the dominance of the computing machinery in medicine develop in the future and whether this development be a blessing or a curse when computers become intelligent in the years ahead. Although this is a serious issue that concerns the very nature of health care, there are only a few people who take it seriously. On the contrary, most physicians and medical officials believe that computers as machines will lack intelligence forever. This presumption will be examined in what follows. To this end, natural human intelligence and artificial machine intelligence are compared with each other in the following two sections:

21.1 Natural Intelligence
21.2 Artificial Intelligence

to inquire into whether genuine artificial intelligence is to be expected in medicine.

21.1 Natural Intelligence

Intelligence is usually understood as the ability to solve specific problems. It is generally assumed that human intelligence is the most developed one on earth. But we should be aware that intelligence is not only an achievement of the brain. Rather, it depends on the whole organism. For example, a student will

© Springer Science+Business Media Dordrecht 2015

K. Sadegh-Zadeh, *Handbook of Analytic Philosophy of Medicine*,
Philosophy and Medicine 119, DOI 10.1007/978-94-017-9579-1_21

not be able to solve a difficult arithmetical problem as good or as fast as her fellow student does if she is suffering from hypoglycemia due to a pancreatic or hepatic malfunction. That means that not only the brain, but also the pancreas, the liver, and other organs causally contribute to the intelligence of an individual. (For the holistic origin of the entire mind, see our palimpsest theory of consciousness and self-consciousness on page 151.)

The common understanding of the term "intelligence" above is indeed a superficial one. The reason is that intelligence is a very complex attribute that cannot be cast in a concise handy concept. Rather, it requires a comprehensive theory of intelligence that we cannot afford here. Actually, a salient fact that is usually ignored precludes a classificatory concept of intelligence, such as "intelligence is the ability to solve specific problems", on the grounds that according to such a concept amoebas and human beings count as members of the same class of intelligent creatures. This is too coarse a taxonomy to shed much light on the nature of intelligence. A comparative concept of the form *being more intelligent than* would be a better choice, whilst a quantitative concept would provoide the most precise, illuminating, and fruitful construct because it would enable us to conceive intelligence in relation to measurable differences between the manifest problem-solving behaviors of individuals. This is just what has been achieved in the history of intelligence research by introducing the concept of IQ, and IQ tests, as quantitative tools for understanding intelligence. It is only through this advanced conception, and its critique, that it has become possible to distinguish several *aspects* of intelligence – such as the ability to remember information, the ability to apply solution strategies to given problems, the ability to imagine, etc. – which are differently distributed among the population. Some cognitive and educational scientists, e.g. (Gardner, 1999, 2011), go so far as to postulate different *types* of intelligence, such as logical-mathematical, verbal/linguistic, spatial/visual, musical, emotional, interpersonal, intrapersonal intelligences and others. The question arises whether it is possible to duplicate some or all of these types of human intelligence in machines.

21.2 Artificial Intelligence (AI)

The commencement of AI research in the 20th century is closely related to the following two millennia-old philosophical questions: (i) How does the human mind work? (ii) Do, or can, non-human beings, be they living things or artifacts, have minds? These two questions have nurtured the dream of creating intelligent artifacts that is as old as human technology.[143] After the construction of electronic computers in the mid-20th century, the British mathematician

[143] It was not until the 17th century, however, that the first promising steps were taken by creating mechanical 'calculating machines'. The prime exemplar of such machines was the *calculating clock* designed in 1620 by the German pioneer Wilhelm Schickard (1592–1635). This device was the mechanical prototype of what

and computer scientist Alan Mathison Turing's speculations on intelligent and thinking machines (Turing 1948, 1950)(see below) incited some young U.S.-American scholars from different disciplines to create such machines. Among them were the computer scientist and mathematician John McCarthy (1921–2011), the computer and cognitive scientist Allen Newell (1927–1992), the social and cognitive scientist Herbert A. Simon (1916–2001), and the cognitive scientist Marvin Minsky (born 1927). For details of the history of AI, see (McCorduck, 1979). John McCarthy, a mathematician at Dartmouth College in Hanover (New Hamshire, USA), organized in collaboration with a few colleagues a two-month international workshop in the summer of 1956 (see McCarthy et al., 1955). In his fundraising proposal to the Rockefeller foundation he said:

> We propose that a two-month, ten-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hamshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves (McCarthy et al., 1955; Russell and Norvig, 2010, 17).

This was the first official usage of the term "artificial intelligence". A new field of research was born at the conference. From the start, but more and more clearly in the course of time, the idea emerged that the aim and scope of AI research was to duplicate human intelligence in machines. In what follows, we shall therefore briefly discuss these topics:

21.2.1 Aims and Scope of AI
21.2.2 Limitations of AI
21.2.3 Is 'AI in Medicine' possible?

to inquire into the prospects of AI in medicine.

21.2.1 Aims and Scope of AI

In 1637, the French philosopher and mathematician René Descartes argued in his *A Discourse on the Method* that a machine can never think (Descartes, 1637). Some 300 years later, the British mathematician, logician and computer scientist Alan Turing (1912–1954) opposed this view in his legendary

---

would come to be called a "computer" in the 20th century. The intermediate inventions were the French mathematician Blaise Pascal's *calculator* (1642), Gottfried Wilhelm Leibniz's *wheel calculator* (1685), and Charles Babbage's *analytical engine* (1837). For details of this evolution, see (Martin, 1992). The emergence of electricity in the 19th century, of mathematical logic at the turn of the 20th century, and of theoretical informatics in the early 20th century enabled the construction of electronic computers as we know them today.

article entitled "Computing machinery and intelligence" in which he put the question "Can machines think?" and answered it, indirectly, in the affirmative (Turing, 1950). His argumentation in support of this affirmation has come to be known as the *Turing Test* that has been a driving force behind the creation of mordern AI research at the above-mentioned Dartmouth workshop shortly after 1950 . In the following sections the Turing Test is outlined and the criteria of intelligence attributed to an 'intelligent machine' are briefly discussed:

- The Turing Test
- Natural language understanding
- Learning
- Knowledgeability
- Reasoning capability
- Vision.

For the sake of clarity, it is worth noting that the term "artificial intelligence" is ambiguous. It has two meanings: (i) a specific one that refers to machine intelligence as an attribute; and (ii) a generic one denoting a research field (AI research). AI research, also called *Computational Intelligence,* is an interdisciplinary field of inquiry. Cooperating disciplines are computer science and technology, logic, mathematics, cognitive science, philosophy of language, philosophy of mind, and many other branches. For details, see (Russell and Norvig, 2010; Kirsh, 1992; Shapiro, 1992).

The Turing Test

To prevent pitfalls of semantics and psychologism, Turing didn't answer the above-mentioned, popular question "Can machines think?" directly by first providing definitions of *machine* and *thinking*. He found the question to be "too meaningless" to deserve discussion (Turing, 1950, 442). Therefore, he preferred answering it indirectly by inventing a game that he called "the imitation game" to show why a machine winning the imitation game must be considered intelligent like human beings. Roughly, his imitation game proceeds as follows (Turing, 1950):

The game is played by three persons, a man (A), a woman (B), and an interrogator (C). A and B are in separate rooms and invisible to C. The interrogator doesn't know which of the two persons is a man and which is the woman. Her task is to identify them from the answers they give to specific questions she puts to them. Such a question could be, for example, "Please tell me the length of your hair!". To prevent trivialization of the game, A and B communicate their answers by a teleprinter. The object of the game for A, the man, is to imitate a woman and thereby to deceive the interrogator, whereas the object for the woman, B, is to help the interrogator about her true gender. If it is the man who receives the above question, he may answer by teletyping "My hair is shingled, and the longest strands are about 22 cm

long", while the woman may add such things as "I am the woman, don't listen to him!" to her answers. But her truthfulness will not help because the man can make similar remarks.

What will happen when a machine (computer) takes the part of A, and a human being the part of B in this game, Turing asked, while the task of the interrogator is to find out which of the two players is a human being and which a machine? It is this variant of the imitation game that is called *the Turing Test*.

The Turing Test provides a scientific test that disambiguates the initial, vague philosophical question "Can machines think?". In a Turing Test, the interrogator asks A and B any questions and receives their answers on a monitor. Again, in this version of the game, A (the computer), gives deceptive answers to arouse the impression that it is a human being. When asked, for example, add 34957 and 70764, it may pause about 30 seconds and give the wrong answer 105621 rather than give a correct answer (= 105721) quickly. Will in this new game with a machine and a human being the interrogator decide wrongly as often as in the previous game with a man and a woman? This question replaces the initial question "Can machines think" or "Can machines be intelligent?". According to Alan Turing, a machine must be considered intelligent, and thus a 'thinking machine', if the interrogator has difficulty deciding which of the two is the human being and which the machine. In this case, it is said that the machine has passed the Turing Test.

The Turing Test is actually a tacit, operational definition (see page 103) of the term "intelligent machine" which says that under the condition required by the test, *a machine is intelligent if and only if it passes the test*. Turing himself claimed in 1950 that in about fifty years it would be possible to program computers with a sufficiently large storage capacity to play the imitation game successfully and achieve the human-level performance in cognitive tasks. We shall see below that in specific domains his prognosis is gradually approximating the truth.

What do computers need in order to achieve the human-level performance? Some abilities that would make them so behave *as if* they really possessed intelligence are discussed in the next five sections. They constitute the core aims of AI research and technology to create intelligent artifacts.

Natural language understanding

Understanding natural language by computers is a very hard task. Research experiences gained over decades have shown that a requisite for natural language understanding is an understanding of the subject matter and context, and this in turn requires encyclopedic knowledge which a computer doesn't possess, however. For example, to understand a context that contains the sentence "that is like carrying coals to Newcastle" requires some knowledge about the city Newcastle upon Tyne in North East England and the history of coal production there since the 16th century.

At first sight it does not seem necessary for a machine to understand natural language if it is to become intelligent. But without natural language understanding a computer will not be able to accomplish so important tasks as translating; communicating with human beings, e.g., history taking in medicine; natural language processing, and so on (see page 61).

Learning

Learning is a basic requirement of intelligent behavior and interaction in changing environments. Knowledge acquisition that was discussed on page 721, is a kind of learning. But it must be viewed as a more or less passive process. Active learning would involve making new experiences that alter the internal states of the system, including its knowledge base. We have encountered the prototype of such active learning in artificial neural networks (ANNs) on page 726, and in all hybrid decision support systems which include ANNs among their components (see pp. 729–731). Systems of this type are able to enhance their knowledge and intelligence autonomously.

Knowledgeability

Knowledgeability means being well-informed and possessing knowledge about some particular domain or domains. For example, the knowledge base of a cardiology expert system renders it knowledgeable about diseases of the cardiovascular system and their treatment. Diagnosis, prognosis, and every other judgment and decision made by a medical AI system depends on its knowledgeability. Learning, be it active or passive, enhances the knowledgeability of an AI system, and thereby, its IQ and judgments.

Reasoning capability

Do higher animals such as dolphins, elephants, and dogs reason? Possibly they do, although their reasoning does not occur by using symbols and sentences like in human beings, but probably sounds, pictures or smells in their memory. It is likely that the strength of the similarity between a past and a present experience plays a central role in their reasoning that therefore may be viewed as a kind of analogical reasoning (see *similaristic reasoning* on page 663). Symbolic *and* similaristic, or analogical, reasoning is a peculiarity of human beings. Both types of reasoning are computationally imitated by machines with the aid of different systems of logic in their programs. Machine reasoning, also called automated reasoning, is a major branch of AI research today.[144]

What is usually called *inference engine* in expert systems (see page 721), is their reasoning capability that may have different styles in different expert

[144] See *Journal of Automated Reasoning* at http://www.springer.com/computer/theoretical+computer+science/journal/10817. Last accessed June 24, 2013.

systems. In any event, the reasoning of an AI system will vary depending on the logic on which its programs and algorithms are based. We have seen in several places in this book that there is a major difference between logics auch as classical predicate logic, paraconsistent logic, and fuzzy logic (see Part IX on pages 891–1120). This brings with it that AI systems exhibit different intelligences if their reasoning uses different, not equivalent logics.

Vision

A sighted AI system, e.g., a gastroenterology expert system associated with an endoscope, will be at an advantage over a blind one because it can see the objects and processes in its domain of expertise and thereby gain information about it, adapt its judgments and advices, etc. But seeing does not only mean sensing. What the optical sensors of an AI system provide about its domain of expertise, say images, must be interpreted and understood by the AI system. *Image understanding* is thus a requisite for an AI system's *vision*.

Image understanding, and thus computer vision, is at least as hard a task as natural language understanding. Theories of physics, geometry and optics, probability theory, fuzzy logic, and other resources alone will not be sufficient to enable such an understanding. And presumably, it will not be possible to technologically recreate the human visual system, i.e., the retinal photoreceptors of the human eye and the yet poorly understood, complicated system of visual information processing in the brain.

21.2.2 Limitations of AI

AI research started as a field of inquiry and technology with the aim of replicating human-level intelligence in machines (see page 735). The preceding five sections demonstrated that for AI research to be successful, it has hard tasks to accomplish. It is even possible that some or all of them will turn out unfeasible. There are indeed vehement critics who argue that AI research will fail because 'artificial intelligence' is impossible. The most prominent one among them is the U.S.-American philosopher John Searle (1980, 1986, 1990, 1992). In the following two sections:

- John Searle's Chinese Room
- Multiple forms of intelligence: AI *vs.* human intelligence

we shall briefly outline his argument to show that the replication of human-level intelligence in machines is an unrealistic goal, whereas AI is something different than human intelligence and definitely achievable.

John Searle's Chinese Room

In capacities such as 'natural language understanding', John Searle considers *understanding* as a mental state. He therefore confuses what is referred to as

intelligence in AI research, with *mind* (see Searle, 1980, p. 417). Regarding the nature and origin of the mind, he takes a radical, biological-naturalistic position and is of the opinion that:

> Mental phenomena are caused by neurophysiological processes in the brain and are themselves features of the brain. To distinguish this view from the many others in the field, I call it "biological naturalism." Mental events and processes are as much part of our biological natural history as digestion, mitosis, meiosis, or enzyme secretion (Searle, 1992, p. 1).

Underscoring the exclusivity of the brain as the sole source of the mind in the world, Searle is in fact a *cerebral naturalist*. According to this worldview, mental states cannot be duplicated in other systems just on the basis of some programs which do not possess the same causal structure and function as biological brains. Even if a machine exhibits the same input-output behavior as a mind, he says, it is void of mental states and understanding like the following *Gedankenexperiment* that has come to be known as Searle's *Chinese Room* (Searle, 1980, 417–418):

> Suppose that I'm locked in a room and given a large batch of Chinese writing. Suppose furthermore (as is indeed the case) that I know no Chinese, either written or spoken, and that I'm not even confident that I could recognize Chinese writing as Chinese writing distinct from, say, Japanese writing or meaningless squiggles. To me, Chinese writing is just so many meaningless squiggles. Now suppose further that after this first batch of Chinese writing I am given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English, and I understand these rules as well as any other native speaker of English. They enable me to correlate one set of formal symbols with another set of formal symbols, and all that 'formal' means here is that I can identify the symbols entirely by their shapes. Now suppose also that I am given a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch with the first two batches, and these rules instruct me how to give back certain Chinese symbols with certain sorts of shapes in response to certain sorts of shapes given me in the third batch. Unknown to me, the people who are giving me all of these symbols call the first batch "a script," they call the second batch a "story," and they call the third batch "questions." Furthermore, they call the symbols I give them back in response to the third batch "answers to the questions," and the set of rules in English that they gave me, they call "the program." Now just to complicate the story a little, imagine that these people also give me stories in English, which I understand, and they then ask me questions in English about these stories, and I give them back answers in English. Suppose also that after a while I get so good at following the instructions for manipulating the Chinese symbols and the programmers get so good at writing the programs that from the external point of view – that is, from the point of view of somebody outside the room in which I am locked – my answers to the questions are absolutely indistinguishable from those of native Chinese speakers. Nobody just looking at my answers can tell that I don't speak a word of Chinese. Let us also suppose that my answers to the English questions are, as they no doubt would be, indistinguishable from those of other native English speakers, for the simple reason

that I am a native English speaker. From the external point of view – from the point of view of someone reading my "answers" – the answers to the Chinese questions and the English questions are equally good. But in the Chinese case, unlike the English case, I produce the answers by manipulating uninterpreted formal symbols. As far as the Chinese is concerned, I simply behave like a computer; I perform computational operations on formally specified elements. For the purposes of the Chinese, I am simply an instantiation of the computer program.

Searle uses his Chinese Room thought experiment to argue that since the man in the Chinese Room, i.e., the human computer, lacks any understanding of what he has accomplished with the aid of the Chinese squiggles, the real-world computers doing similar things also lack any understanding of what they are doing. Therefore, they cannot have mental states. To substantiate this reasoning, he draws from his thought experiment the following three axioms (Searle, 1990, 26–27):

1. Computer programs are formal (syntactic),
2. Human minds have mental contents (semantics),
3. Syntax by itself is neither constitutive of nor sufficient for semantics

to conclude that computer "Programs are neither constitutive of nor sufficient for minds" (ibid., p. 27).

One should be aware that Searle's thought experiment and reasoning above do not provide a cogent argument against the possibility of creating intelligent artifacts. For the Chinese Room *as a whole*, not Searle himself as a part in the whole, is obviously an intelligent system. It even has passed the Turing Test as Searle himself admits that "my answers to the questions are absolutely indistinguishable from those of native Chinese speakers" (see above). The definition of the Turing Test on page 736 implies that passing the Turing Test is sufficient for the ascription of intelligence to a machine. (For criticisms of the Chinese Room from other perspectives, see Preston and Bishop, 2002.)

Multiple types of intelligence: AI *vs.* human intelligence

The conclusion of the last section means that artificial *intelligence* should not be confused with artificial *mind*. Intelligence is an objective cognitive ability and not a subjective mental state of a system (see page 144). Independent of whether there will, or will never, exist a machine *mind,* the question of whether machine *intelligence* is possible, is not a subject of speculative worldview. To decide the question, introduce a machine IQ test, a MIQ test so to speak, and just use it. A prerequisite for doing so is some clarity about what is to be understood by the term "artificial intelligence" or "machine intelligence". It is a mistake to equate this type of intelligence with human-like intelligence. On the one hand, it is unrealistic to expect an intelligent machine to exhibit all aspects of human intelligence, e.g., subjective feeling of understanding. On the other hand, an intelligent machine's intelligence will have aspects that a human being will certainly lack, e.g., extreme knowledgeability and tremendous

multi-logical reasoning capability. It is therefore advisable to view machine intelligence and human intelligence as two different types of intelligence which are not competitive, but complementary. Thus, a final caveat is: Don't believe that AI is, or ought to be, the replication of human intelligence in machines! Anthropomorphism as well as anthropocentrism should be avoided.

21.2.3 Is 'AI in Medicine' possible?

Skepticism about the possibility of AI is fostered primarily in the humanities where anthropocentrism is at home and man is still the apex of creation. See, for example, (Dreyfus and Dreyfus, 1972; 1986; 1992; Searle, 1980, 1986, 1990, 1992). In medicine, however, critical attitude toward the computing machinery is steadily diminishing as new generations of students and physicians are growing up with computers. There are also other factors, such as commercial and educational ones, expediting the proliferation of computers and clinical informatics with AI in medicine. In the present context, the question arises whether 'artificial intelligence' in medicine is possible at all. Although AI techniques are applied in medical research as well as practice, we will consider here the latter area.

Physicians are usually viewed as the best diagnostic-therapeutic decision-makers superior to computers because it is generally believed that they possess *intuition* and "can catch on to a joke" (see Wartofsky on page 335). Computers are viewed as mere metallic and silicon devices void of any intelligence, reason, and intuition. But contrary to this widespread rumor, what is called *intuition* in a human being, is an emergent, subjective feeling that corresponds to subconscious neural processes (for the term "emergent", see pages 131–133). Intuition is no capacity, and therefore both causally inert and superfluous for clinical judgment. Ever since its inception in the 1970s, medical AI research has made tremendous progress. It would be a mistake today to overlook the following facts:

In limited domains such as the diagnostics or treatment of specific diseases and in confined medical specialties such as heart diseases or rheumatology, clinical decision support systems are at least as intelligent as physicians. As AI systems, they achieve a remarkable performance that even exceeds the physician performance. For instance, the evaluation of an early decision support system for use in the diagnostics of acute abdominal pain showed an overall diagnostic accuracy of 91.8%, whereas the clinicians' diagnoses were correct in only 65–79.6% of the 304 patients (de Dombal et al., 1972a–b). Another such example is the clinical decision support system DXplain briefly described on page 718. In a comparative study of four computer-based diagnostic systems, it achieved an overall performance of 91% (Berner et al., 1994, p. 1794). The performance analyses of clinical decision support systems report, in general, a diagnostic accuracy of 50%–95%.

That *some* current AI systems in clinical decision-making are more accurate than physicians, may be correctly interpreted thus: (i) Such systems are

obviously able to pass the Turing Test (see page 736) in that in 'diagnostic games' they outrun the physicians' clinical reasoning capability; (ii) they are even more intelligent than physicians. This result can be considered a proof of the possibility of creating AI in medicine.

We should be aware that AI in medicine is still in its infancy. There is no doubt that by the end of this century at the very latest, it will have progressed so far that in clinical practice clinical judgment will constitute its domain of competence. Hospital information systems with integrated intelligent clinical decision support systems (see page 731) will thus completely automate clinical judgment. They will act as clinical process control systems making diagnostic and therapeutic decisions and using the health personnel as their mobile peripherals (see page 337).

It seems necessary to devote some thought to the causes and consequences of this fascinating development. Maybe the present reader will feel the need to evaluate the imminent development upon realizing that the medical knowledge base required by the automated clinical judgment alluded to above will also be produced ('engineered') by AI itself, specifically by AI in medical research (see Section 13.3.3 on page 566).

21.3 Summary

Natural intelligence and artificial intelligence (AI) are briefly compared to examine whether (i) AI is a duplication of human intelligence in machines and (ii) AI in medicine is possible at all. To this end, the concept of AI and the aims and scope of AI research are analyzed. As a basic notion of AI philosophy the so-called Turing Test is introduced. The famous argument against the possibility of AI put forward by John Searle, his so-called Chinese Room, is briefly discussed to show that the argument is self-defeating as it clearly demonstrates the possibility of AI. It is argued that AI is a new type of intelligence and does not represent a duplication of human intelligence in machines. AI in medicine is not only possible, but it exists already. Key pieces of evidence are a number of clinical decision support systems of high performance which have clearly passed the Turing Test already.

Theoretical Medicine and Bioethics
https://doi.org/10.1007/s11017-018-9442-3

Against the iDoctor: why artificial intelligence should not replace physician judgment

Kyle E. Karches[1]

© Springer Nature B.V. 2018

Abstract

Experts in medical informatics have argued for the incorporation of ever more machine-learning algorithms into medical care. As artificial intelligence (AI) research advances, such technologies raise the possibility of an "iDoctor," a machine theoretically capable of replacing the judgment of primary care physicians. In this article, I draw on Martin Heidegger's critique of technology to show how an algorithmic approach to medicine distorts the physician–patient relationship. Among other problems, AI cannot adapt guidelines according to the individual patient's needs. In response to the objection that AI could develop this capacity, I use Hubert Dreyfus's analysis of AI to argue that attention to the needs of each patient requires the physician to attune his or her perception to the patient's history and physical exam, an ability that seems uniquely human. Human physician judgment will remain better suited to the practice of primary care despite anticipated advances in AI technology.

Keywords Technology · Artificial intelligence · Electronic health records · Physician judgment

Introduction

Ours is an age of ambivalence about technology. On the one hand, we find ourselves enamored of its promise to extend our abilities and make our lives easier. Consider a recent viewpoint article by Ravi Parikh et al. in the *Journal of the American Medical Association* arguing for the incorporation into clinical care of "predictive analytics," algorithms that use historical information to predict future outcomes [1]. The authors cite examples such as Amazon's product recommendation system for online shopping, and they claim that such "sophisticated machine learning algorithms"

✉ Kyle E. Karches
kyle.karches@health.slu.edu

[1] Department of Internal Medicine, Saint Louis University Hospital, 3635 Vista Ave, Saint Louis, MO 63110, USA

Published online: 10 July 2018

could analyze "big data" from electronic health records (EHRs) to predict patient risk and direct resources accordingly [1, p. 651]. For example, Parkland Memorial Hospital in Dallas already uses such an algorithm to identify patients at high risk for hospital readmission. Although the authors note concerns that these systems could threaten patient privacy and diminish the role of the physician's judgment, they claim that "algorithms routinely outperform practitioners' clinical intuition" and may help reduce the cost of care in the United States [1, p. 652]. If they are indeed correct about the power of algorithms, one might wonder whether or not advances in artificial intelligence could eventually replace physicians altogether, just as corporations such as Google promise to produce driver-less cars.

On the other hand, there is a nagging sense that the advance of technology effaces something important. A recent article for the *New York Review of Books* surveys the emerging literature about how smartphones are changing the way young people interact with each other and with the world [2]. For example, a psychologist concludes from a series of hundreds of interviews at schools that smartphones corrode the capacities for empathy and conversation [3], creating "students who don't make eye contact or respond to body language, who have trouble listening and talking to teachers, and can't see things from another's point of view, recognize when they've hurt someone, or form friendships based on trust" [2]. She cites the example of a family that holds arguments on Gchat because the "value proposition" of face-to-face conflict is low [3, p. 127]. The online chat tool simply makes them more "productive." In fact, designers of smartphone applications intend this colonization of daily life. Many of these designers studied how to manipulate human psychology at Stanford's Persuasive Technology Lab, learning ways to produce behavioral loops that keep consumers returning to applications time and again. Yet the author of the review resolves that the proposed solutions to these concerns, such as "more thoughtful apps" or exhortations to "reclaim conversation," are "wildly inadequate" [2]. The problems with technology seem hopelessly intractable.

In this paper, I critique the notion that primary care physicians are able to be replaced by an artificially intelligent "iDoctor."[1] I develop two related lines of argument. First, I explain the critique of technology developed by Martin Heidegger and Albert Borgmann and use it to bring into focus the potential deleterious consequences of replacing physicians with artificial intelligence (AI) machines. I then draw on Hubert Dreyfus's analysis of AI [4], which in turn relies on Heidegger's epistemology and the work of Michael Polanyi [5, 6], to argue that the judgment of human physicians is better suited to the practice of medicine than is AI. As an example of what Borgmann calls a focal practice [7], primary care would inevitably be distorted by the loss of the human physician.

[1] For ease of reference, I use the term *primary care physician* throughout this paper. However, in so doing I do not mean to exclude other providers involved in primary care, such as nurse practitioners and physician assistants.

Heidegger's and Borgmann's critique of technology

Heidegger's critique of technology derives from his more general critique of metaphysics as a succession of ontotheologies [8–10]. Heidegger claims that, starting with Plato, Western metaphysics divided the question of Being into two parts. The first is ontological and asks what makes an entity an entity, an inquiry into the essence of entities and into what all entities share in common. The second is what Heidegger calls theological and asks what the source or ground of being is, an inquiry into the existence of entities and the question of which entity is the highest being. Far from a mere intellectual exercise, metaphysics creates an ontotheology that shapes the way individuals understand Being itself—determining their basic presuppositions about everything in existence, including those about themselves and, as regards physicians, their patients. Heidegger thus builds on the Kantian notion that people participate in making their worlds intelligible [10, p. 53]. The way in which Being appears to them depends upon their metaphysical assumptions, the ontotheology with which they engage the world.

Heidegger interprets the history of Western thought as a succession of ontotheologies, ending with Nietzsche. Whereas Nietzsche thought of himself as antimetaphysical, Heidegger perceives an ontology in Nietzsche's will-to-power—which Nietzsche believed to be fundamental to all entities—and a corresponding theology in the "eternal return" of the will-to-power fully actualized, the pinnacle that any entity can achieve [10, p. 21]. Heidegger argues that the ontotheology of our late modern epoch is fundamentally Nietzschean. Being itself appears to us in the form of power relationships, as the constant interaction between competing forces with no inherent nature or purpose other than self-perpetuation through the actualization of their will-to-power. We perceive things in the world as intrinsically meaningless, awaiting the operation of human will-to-power to fashion their meaning.

According to Heidegger, modern technology is part of this Nietzschean ontotheology. He argued that "the essence of technology is by no means anything technological" [5, p. 311], but rather a way of engaging the world called *Gestell*, or "enframing" [5, p. 322]. The Nietzschean notion that things have no inherent meaning allows technology to conceive of everything in the world as *Bestand* or "standing-reserve," that is, as mere raw materials or resources awaiting the imposition of order by the human will [5, p. 322]. Thus, when we moderns look at a tree, we see it in terms of the amount of lumber that it yields rather than on its own terms. Technology is a background assumption about the world that shapes the way the world appears to us.

Technological enframing leads us to extract things from their context within the world so that, for example, we view natural entities as sources for abstract gains like energy, to be harvested and stored. Indeed, the natural object is eclipsed from our view, as the energy becomes more real than its source. We reduce the qualitative to the quantitative, to mere information or data, until it is only the quantifiable that matters at all. Most perniciously, according to Heidegger, in late modernity we have turned the frame on ourselves, such that we become "human resources" awaiting optimization for maximally flexible use [10, p. 60]. This technological

understanding of human beings makes them, like everything else in nature, available for manipulation by external forces, such as the market economy.

Within the technological enframing, the primary criterion for evaluation of anything is efficiency. In the debate over clean energy, for example, one side claims that wind power is better than coal because it more efficiently meets our desire for minimally toxic energy, whereas the other side claims superiority for coal because it more efficiently meets our desire for cheap energy. Yet both sides look at objects in the world according to *Gestell*, not as things with their own nature but rather as resources for us to subject to our own desires. Heidegger calls this stance toward things in nature a *challenging-forth*, a demand that they conform to our wishes, the efficient fulfillment of which determines their moral worth [5, p. 320]. We even subject ourselves to this analysis, explaining why we so readily accept the replacement of human labor by machines, despite the human consequences: if a machine can accomplish the task more efficiently, then so be it. In a world ordered by the will-to-power, human beings enjoy no unique status but rather become one intrinsically meaningless force among others [11, p. 184].

For Heidegger, the primary problem with *Gestell* is that, like all ontotheologies, it reveals certain aspects of Being but obscures others. The smartphone applications cited in the introduction may make us more efficient, for example, but in using them we miss other important aspects of human interaction. As technology advances, we even lose the ability to perceive some of nature's properties, particularly its qualitative properties. Heidegger contrasts the challenging-forth of technology with the *bringing-forth* of traditional arts and crafts, which are ontological "openings" that allow things to reveal what they truly are in their fullness and to order our knowledge of them, prior to the will-to-power's operation upon them [11, p. 184]. For Heidegger, the solution to the problem of technology is ultimately this way of "letting beings be" [12], rather than imposing an ontotheology upon it at the outset. The only way to gain a free relation to technology is to understand it for what it is, an enframing that discloses some aspects of the world in certain ways and conceals other aspects of it in other ways. If we apprehend technology along these lines, perhaps we can step outside *Gestell* so that it no longer dominates us, and thus our epoch's technological ontotheology would give way to another.

Borgmann extends this Heideggerian critique, suggesting a concrete way in which to overcome the technological enframing. He describes the pattern of modern technology in an analogous way, as the *device paradigm*, which reduces the full significance of some part of human life to its essential function and then realizes that function as efficiently as possible [7, p. 40–48]. Within this paradigm, the only relevant criteria for moral evaluation of technology are instantaneity, ubiquity, safety, and ease of use. Devices make commodities easily available for consumption while typically concealing their inner workings, thus disburdening human beings of the labor that used to be required to realize the goods in question. Although they make human life more convenient in some ways, they also degrade the importance of skilled human labor, compressing human excellence into the small group of elite engineers responsible for technological design. The resultant attenuation of skill in the population then necessitates the development of ever more technology to compensate. Devices also disengage human beings from the real world and from each

other, rendering unnecessary the bodily expertise and caring attentiveness that characterize pre-technological practices.

To illustrate the device paradigm, Borgmann contrasts the pre-modern hearth with the heating systems of modern houses. In addition to providing warmth, the hearth was the center of work and leisure for a family's home. It required bodily engagement and the development of skills to procure wood and build the fire, necessitating the division of labor among members of the household. Thus, the hearth was inseparable from the practices that shaped families' experiences of the world and developed their capacities. Modern heating systems, by contrast, dissociate the good of warmth from this rich context, commoditizing it for easy consumption at the flip of a switch. Such systems conceal their machinery, making them inaccessible and incomprehensible to their users, so that a specialist is needed when the machinery fails. Borgmann claims that this disengagement from the material world occasioned by technology deprives human beings of opportunities to pursue excellence within a community dedicated to the achievement of certain goods. Ironically, although technology promises to augment our freedom and our abilities, it has instead undermined these human capacities.

Borgmann argues for a reform of technology centered upon what he calls "focal things" and "focal practices" [7, p. 196–219]. *Focal things*, such as the wilderness and the family meal, have dignity and greatness in their own right and therefore ennoble human life [7, p. 217–220]. They engage human capacities fully, in part because they cannot be possessed or controlled. *Focal practices*, then, are the socially established activities dedicated to focal things. They build up habits of engagement with the real world, fostering discipline and excellence of mind and body. According to Borgmann, focal things and practices are fragile, ever in danger of being undermined by technology, which tends to make the sustained effort required for focal practices unnecessary or unappealing. Yet they can also provide the orienting force for a reform of technology, because they give people reasons to use technology more selectively as a means of promoting the kind of human excellence that can be achieved only through focal practices, thereby restraining the device paradigm. For example, instead of using a microwave to heat up frozen food for a family dinner, thus bypassing human effort entirely, one might decide to use only kitchen gadgets that help to extend one's cooking skills, thus preparing a truly better meal. Borgmann calls for a renewal of the political discourse that attests to the importance of focal things in human life, revealing them in their greatness, much like poetry. Such discourse may help to identify the points at which technologies threaten the achievement of authentic human excellence within focal practices.

Technology in primary care

With this conceptual framework in place, one can analyze the way in which primary care physicians currently engage two different technologies: the stethoscope and the electronic health record (EHR). Although the stethoscope is certainly a form of technology, it also fits Heidegger's description of what he calls a simple tool. In a phenomenological account of the use of such tools, Heidegger points out that they

recede from our attention as they are used. When one uses a hammer, for example, one focuses not on the hammer but rather on the nail [13, p. 7]. Similarly, the physician using a stethoscope directs his or her attention to the sounds of the patient's bodily functions. The withdrawal of the tool's own presence allows the human operator to assimilate it and inhabit it, such that it becomes an extension of his or her body. Thus, as opposed to other technologies, simple tools promote direct human contact with the world. The stethoscope brings the physician closer to the patient's authentic body and helps the physician attune his or her senses to it.

The stethoscope therefore exemplifies Borgmann's ideal of technology, casting its use as a means of promoting the goods attainable within focal practices. At its best, primary care is a focal practice dedicated to the care of the patient, a focal thing of ultimate concern. Success in this endeavor requires the physician to attune his or her senses and intuitions to the patient's body and its needs. Because it acts as an extension of the physician's own human senses, the stethoscope facilitates this process. It directs the physician's attention to the patient's body as it really is, allowing the physician to conform his or her judgment to this bodily reality. Thus, the stethoscope helps the physician "bring forth" health from the patient's body, in Heideggerian terms.

The EHR, by contrast, tends to distance the physician from the patient's living body. Often the physician has access to the patient's EHR before he or she meets the patient. Before the physician walks into the exam room, the patient's vital signs appear in the record and the *problem list*, a list of the patient's symptoms and medical conditions, is generated. Thus, the EHR draws the physician's attention away from the patient's actual body toward a collection of facts about the patient's body, to the point that physicians now spend more time at the computer than at the bedside [14]. The lack of attention to the physical examination of patients has raised concern among medical educators that the new generation of physicians will no longer be able to perform an adequate exam. Reliance on technology leaves physicians less equipped to perceive aspects of patient care that cannot be captured technologically.

Further developments in EHR technology, such as so-called best practice advisories (BPAs), tend to treat these facts about the body as things of ultimate concern. In the Epic EHR system that I use in my practice, for example, every patient encounter triggers BPAs that remind the physician to meet all of the federally mandated quality measures appropriate for the patient's age and sex, such as cancer and cholesterol screening. Just as technological enframing leads one to view a tree in terms of the amount of lumber it can yield, so does the EHR encourage physicians to view patients' bodies in terms of their most basic characteristics. It especially focuses physicians' attention on those quantitative data such as blood pressure and cholesterol that can be manipulated with medications. As a set of algorithms, it can neither account for the individual patient's preferences and circumstances nor leave room for the physician to interpret the guidelines. Instead of drawing the physician closer to the patient's body, in all its uniqueness, the EHR treats it abstractly, challenging it forth from its context and constituting it as a set of facts over which the physician can exert control.

The EHR also exposes patient care to manipulation by forces external to the doctor–patient relationship, such as the state or insurance companies. Because the

record is digital, unlike older paper charts, third parties have easy access to patient information. Insurers, including the federal government, have begun to extract data from EHRs to provide primary care physicians with reports of the rates at which their patients meet certain quality measures, such as blood pressure control or colonoscopy at age fifty. In the near future, they will tie physician reimbursement to these benchmarks, encouraging physicians to complete them for each patient as soon as possible. Thus, the EHR draws the physician's attention toward an even higher level of abstraction, away from the health of the individual patient's body and toward the health of the body politic. It turns information about patients' bodies into the means by which forces such as the state can exert power over them, achieving the good of health with maximal efficiency.

Advocates of EHRs insist that algorithms like BPAs improve preventative care, pointing out that they increase rates of completion of screening measures [15]. On this model of primary care, the best physician is the one who performs all screening measures on each patient at every visit. Yet such a physician may neglect to explain the purpose of these tests or to ask patients whether or not they want the tests done. He or she may also focus less attention on problems that patients consider more important. In other words, this model of primary care leaves no space for an alternative vision of the good primary care physician as one who applies guidelines to each patient's unique circumstances, taking the patient's own preferences into account. Such a physician may even choose temporarily to ignore the guidelines altogether in order to address a patient's main concerns. Advocates of EHRs may argue that physicians using EHRs can still maintain this standard of excellence, but when EHRs tie compensation to completion of screening measures, the busy primary care physician is unlikely to resist the temptation to focus on them at the expense of other matters.

The EHR and its economic incentives thus threaten to turn primary care physicians themselves into resources to be optimized for efficiency. Physicians in the era of high throughput find themselves under administrative pressure to see more patients in less time using fewer resources and to devote less time to uncompensated activities such as teaching students. The BPAs are intended to facilitate higher throughput by simplifying patient care, providing physicians with a checklist to accomplish at each visit. Yet physicians subjected to such expectations of efficiency have begun to feel as though they provide "care on a production line" [16]. Due to these constraints, they cannot offer the individualized attention and care that each patient expects. It is perhaps little wonder, then, that physician burnout has become a "public health crisis" [17].

As this analysis indicates, medical technologies such as EHRs are not merely neutral tools, to be used as a means to whatever ends their human operators designate. Rather, as Peter-Paul Verbeek has argued, medical technology mediates our experience of the human body [13]. It constitutes the way in which the patient's body appears to the physician. Whereas the stethoscope draws the physician closer to the patient's actual physical body, the EHR constitutes the patient as a body of facts that can be made into inputs for algorithmic reasoning. Currently, this logic remains confined in certain ways. My practice's Epic EHR system has BPAs only for certain preventative care measures and chronic conditions, and even though the EHR mediates the doctor–patient relationship in powerful ways, physician and patient still

encounter each other face-to-face in the office. Yet the replacement of primary care physicians with artificial intelligence would extend the EHR's technological frame to cover all aspects of patient care. As I show in the following sections, such an extension would have far-reaching effects on the doctor–patient relationship.

AI and the doctor–patient relationship

The EHR, with all its algorithms and dictates for patient care, opens up the possibility of the iDoctor, an artificially intelligent machine designed to deliver primary care. This introduction of AI to replace physician judgment would subject the already strained doctor–patient relationship even further to technological enframing. On the patient side, an AI computer would require a mechanical view of human beings as entities comprehensible according to laws that can be programmed into the machine. It would thus isolate the human being from its rich context in the world and deconstruct the human body into its component systems. This tendency toward deconstruction is already latent in modern medicine—as evident in the systems-based reporting of the physical exam and the EHR's collation of a problem list that presents the patient as a concatenation of his or her medical conditions. The context of the patient's life is included in a so-called social history, but usually only insofar as it informs diagnosis and treatment. While many primary care physicians pride themselves on maintaining a focus on the whole patient, AI would extend and systematize the trend toward deconstruction, making its own rationalized representation of human beings more real, in a way, than actual patients. As Heidegger argues, the iDoctor's reduction of human beings to a set of medical concepts necessarily fails to understand life as it truly is, in its fullness.

On the physician side, the aforecited *Journal of the American Medical Association* piece by Parikh et al. [1] offers an example of the morality that accompanies the technological enframing analyzed by Heidegger. Its highest good is efficient diagnosis and therapy, and its means include cost–benefit analyses and the creation of protocols for care. If AI can execute these directives better than human physicians can, then so much the worse for the humans. Concerns about patient safety and privacy are apparently outweighed by the prospect of ever better technology. As Parikh et al. note, detractors of EHRs initially expressed such apprehension, but advances in technology neutralized their criticisms [1]. Every technological problem seems to have a technological answer.

Yet many primary care physicians express frustration at the introduction of technologies like EHRs precisely because they undermine the relationship whereby physician and patient determine together, in conversation, how best to proceed. The sort of algorithms required for AI presume that medical judgment can be made abstractly, without regard for each patient's particular circumstances, for any such consideration would come at the cost of reduced efficiency. The substitution of such techniques for physician judgment would thus effect the ultimate transformation of primary care into a production line. As manifest in the introduction of best practice advisories into the EHR, these algorithms conceive of patients as fungible. For example, they cast every person who turns fifty years old as a patient-in-need-of-a-colonoscopy,

regardless of the patient's circumstances and preferences. The purpose of these algorithms is to convert every such individual into a patient-who-has-had-a-colonoscopy, much like the worker on an assembly line who performs the same action on the same product over and over again.

Whereas best practice advisories currently impose machine algorithms on preventative care, the replacement of physicians with AI would extend the algorithmic model to cover all aspects of primary care. All patients with similar symptoms would be classified together for analysis, and the diagnosis and therapy considered best for such patients would be provided. For example, AI would tend to treat all patients with knee osteoarthritis the same way, according to guidelines, heedless of the differences between particular patients that, in everyday clinical medicine, often determine the best course of treatment. This sort of medicine might maximize efficiency, but the best word to describe it is *mindless*: it delegates no role to the human mind's capacities for creativity and adaptability in accounting for patients' unique circumstances and preferences. Artificially intelligent medicine would replace this type of flexible knowledge with the uniform logic of the assembly line.

The introduction of AI into medicine would thus commodify patient care according to Borgmann's device paradigm. Much like modern heating systems, AI would replace an organic context of skillful human interaction with a machine, whose inner workings remain obscure to anyone without the requisite technological knowledge. Instead of interacting with a fellow human being, the patient would encounter a device created and maintained by experts who need not be present in the medical context at all. Such a "relationship" would be purely functional, focused solely on the provision of diagnosis and therapy as commodities for patient consumption. As with all commodities, the primary moral imperatives for this sort of care would be safety, instantaneity, and ease.

This disburdening effect of AI undermines physicians' capacity for true excellence. The reliance on machine algorithms would absolve physicians of the responsibility to develop their own medical judgment, a form of practical wisdom. This virtue arises in part from individual experience and personal habits of excellence, such as reading the medical literature and learning new skills. But it also arises through collaborative enterprise: it cannot be sustained apart from a community of practitioners mutually dedicated to achieving the goods of medicine or an educational system designed to cultivate good judgment. In other words, as Borgmann shows, excellence in human endeavors like primary care requires a practice. Alasdair MacIntyre has argued that human practices not only promote excellence in the achievement of certain goods but also foster moral virtues such as justice, courage, and honesty [18]. The use of AI in primary care may eventually prove to be safer and more efficient than the use of human physicians, but such use would render unnecessary the formation of communities of good practice that give rise to intellectual and moral virtues. It would also deprive patients of their own role as teachers. Because primary care requires practical knowledge, physicians can develop skills and virtues only in relationship with the patients for whom they take responsibility.

One might object here that AI's potential benefits to medicine outweigh its risk to the doctor–patient relationship. After all, many patients likely see physicians primarily to obtain diagnosis and treatment as efficiently as possible.

As the technology develops, it may turn out that AI machines are simply better at this task than human physicians are. Yet, according to Heidegger's analysis, this objection rests upon a notion of the goods of primary care that is already enframed. For example, the argument that BPAs improve preventative care is quantitative, based on statistics. The BPAs make physicians order colonoscopies earlier and more often than they would otherwise, and the increased rate of such orders is held to be good. As Heidegger points out, however, this enframing reveals certain aspects of Being while concealing others. It is true that, according to evidence from observational studies of the population of all patients aged fifty or older, screening colonoscopy may prevent some deaths from colorectal cancer [19]. Yet a physician who focuses on the provision of such screening measures at the population level interprets his or her individual patients' bodies according to such facts in order to facilitate efficient provision of services. He or she sees the patient as a member of a cohort of individuals that share only their age in common and treats that status as quasi-pathological, requiring a medical procedure as soon as possible. Such a physician might attend to the particular patient's circumstances, but he or she would feel pressure to use this information to convince the patient to undergo the procedure. After all, the physician must ensure that the rate of colonoscopy screening among his or her patients remains as high as possible, so as to maximize the diagnosis of early colon cancer in the population.

Heidegger's and Borgmann's critiques of technology suggest the possibility of an alternative conception of primary care as a focal practice. Instead of constituting the patient's body as a set of facts available for efficient manipulation, primary care physicians may let beings be, so to speak. The physician may attune his or her attention to the patient's body, allowing the body, in all its uniqueness, to call forth the proper response. The focus is primarily on the patient's specific wants and needs, which may not always be amenable to efficient management but rather require a long-term relationship between physician and patient. In most cases, the physician provides care consistent with generalized guidelines, but he or she remains willing to set the guidelines aside when appropriate. This sort of primary care requires the ability to attend to each patient as a unique individual rather than as part of a cohort whose members can all be managed in kind. It resists commodification because it cannot be easily separated from its context within the doctor–patient relationship.

Proponents of AI would likely contend that, eventually, AI machines will become capable of the thought processes underlying this type of primary care practice. Programmers simply need to develop more sophisticated and adaptable algorithms to match the capacities of the human mind. After all, engineers have devised technological solutions to many other problems that once seemed insurmountable. Yet just as Heidegger's critique of technology reveals the iDoctor's potential to dehumanize medicine, so does his epistemology provide a response to this objection. As I show in the next part of this paper, Dreyfus's Heideggerian analysis of AI gives reasons to believe that AI may never be able to attune to human beings in the way that good primary care requires. It may be the case that human thought is not simply more complex than machine thought, but rather wholly different from the type of intelligence that machines can achieve.

Dreyfus's critique of AI

In *What Computers Still Can't Do*, Dreyfus calls attention to the often unspoken rationalist theory of mind underlying AI [4]. Rationalists like René Descartes think all mental knowledge consists of representations containing the fixed, context-free, abstract features of a domain. On this view, such representations can be expressed in propositional form, and indeed the ability to represent knowledge of a domain discursively explains that domain's intelligibility to the human mind. As Dreyfus puts it, this type of representational rationalism "assumes that underlying everyday understanding is a system of implicit beliefs" [4, p. xvii]. Even common-sense know-how is really a type of knowing-that. Here, *know-how* refers to the type of practical knowledge that enables day-to-day interactions with the world and allows one to acquire skills, whereas *knowing-that* refers to the theoretical knowledge that can be adequately contained in the form of propositions.

Similarly, proponents of AI assume that all knowledge can, in principle, be represented as propositions in a vast database, fed to a computer as inputs. Furthermore, they assume that the human mind's manipulation of this knowledge can be reduced to a set of formal rules, which can also be programmed into the computer. The challenge for AI is to generalize background knowledge, such as common sense, away from its context, in much the same way that ontology seeks to describe context-free entities as a foundation for philosophy. Yet as Dreyfus shows in his extensive review of the history of AI, attempts to overcome this challenge have repeatedly failed. Versions of AI produced to date have had difficulty answering certain simple questions rendered in ordinary English, for example. Dreyfus suggests that these failures stem from an inability to capture our *background knowledge*, the "understanding we normally take for granted," in propositional terms [4, p. xix].

Dreyfus turns to Heidegger's epistemology for an alternative to the rationalist theory of mind. As a phenomenological account of skill acquisition shows, human know-how cannot simply be represented as knowing-that. A novice learning a skill such as chess must indeed begin by "learning and applying rules for manipulating context-free elements," but as one gains experience, one learns to find meaning in "context-dependent characteristics such as unbalanced pawn structure" and to see these elements in terms of one's own goals [4, p. xii]. Over time, one leaves the rules behind altogether, having begun to see immediately what to do. The chess master's past experience shapes even her perception of the present moment in the game, such that a response comes to mind without needing to consult a set of general rules. Indeed, the best players are those who creatively transcend the rules. If one does not experience the mastery of a skill as the application of the rules learned through its acquisition, Dreyfus reasons [4], there is no basis for assuming that these rules still play any role at all. The master may sometimes, but not always, be able to explain post hoc the process by which she arrives at a correct response; but, even so, the rule derived from this explanation is not the correct response's cause, as AI proponents assume.

Furthermore, on a rationalist model, for a computer to respond appropriately to a specific situation, it needs to categorize the situation, use rules to search its database for additional rules that could be relevant, and then deduce a conclusion about how

to proceed. The more information about new situations that the computer receives, the more difficult such a process becomes. Dreyfus points out that this very problem has confounded attempts to increase AI's scale, for as the number of inputs increases, the system takes longer to retrieve information [4, p. xxi]. By contrast, as the example of the chess master shows, when a human being gains experience, he or she responds more easily, retrieving the relevant information even faster. This observation indicates that humans rely upon a very different type of information storage than that presumed in the AI rationalist model. When one gains enough experience to become an expert, one's knowledge comes to structure one's very perception such that one "directly experiences which events and things are relevant and how they are relevant" [4, p. xxviii]. As Heidegger would say, objects appear to the expert "not in isolation and with context-free properties but as things that solicit responses by their significance" [4, p. xxviii].

Know-how, then, seems to be built not on rules but on a type of pattern recognition that can be extended to new situations, as well as an immediate, intuitive perception of proper response. Inspired by Polanyi's account of tacit knowledge [6], Dreyfus draws attention to several features of background understanding, which underlie human know-how and would be difficult for AI to replicate [4, pp. xxvii–xxix]. First, as the example of chess indicates, this type of knowledge seems to require vast experience of typical cases. The reason why children "find it fascinating to play with blocks and water day after day for years," Dreyfus suggests, is that they are "learning to discriminate the sorts of typical situations they will have to cope with in their everyday activities" [4, p. xxvii]. Any attempt to capture this knowledge in terms of propositions for a computer program would be so incredibly complex and subject to so many exceptions that its success seems nearly impossible. Yet human children possess and act upon such knowledge intuitively.

Second, much of human background knowledge requires socialization. As Heidegger and Maurice Merleau-Ponty point out, in everyday know-how, what appears as a relevant fact depends on social skills. To demonstrate this point, Dreyfus draws on Bourdieu's description of a gift-giving culture [20]. Members of this culture do not appeal to rules to deduce what to do or how to act but rather simply react in appropriate circumstances with an appropriate gift. The rules need be explained only to an outsider who lacks the requisite social skill. Hence knowledge of gift-giving "is not a bit of factual knowledge, separate from the skill or know-how for giving one" [4, p. xxiii]. Another example is reasoning by analogy or metaphor. One cannot understand a metaphor such as "Sally is a block of ice" representationally by listing the qualities of Sally and ice and then comparing them [21]. Rather, as Heidegger and Merleau-Ponty contend, the social and cultural context in which such devices are deployed allows the relevant association to make itself known.

Third, and most important, know-how often requires the type of knowledge acquired by sensation, emotion, and imagination. One example of the sort of simple English sentence that can stymie AI is: "Mary saw a dog in the window; she wanted it." Computer programs have had difficulty determining whether "it" refers to the window or the dog [22, p. 200]. Dreyfus suggests that interpreting the contextual antecedent of the pronoun in this sentence appeals to "our ability to imagine how we would feel in the situation, rather than requiring us to consult *facts* about dogs and

windows" [4, p. xix]. It also relies on our experience of dogs as objects of desire, an experience gained through socialization. If the latter part of the sentence were instead "she pressed her nose up against it" [22, p. 200], then interpreting the contextual antecedent of "it" would involve appealing to the sensory experience of contact with a window. Such examples require imagination, and recollection of certain experiences, to organize the knowledge necessary for understanding the sentence, and thus they demonstrate the important role of the body in such understanding.

Dreyfus concludes from these observations that our experience and knowledge of the world take their structure from pre-conceptual background knowledge. Human beings approach the world with certain expectations, and meaning arises from the complex equilibrium between these expectations and information from the world. When the expectations that frame one's perceptions are wrong, the incoming data do not make sense, and a new hypothesis is required. When these expectations do succeed, however, one finds oneself able to cope with the world well enough—a situation that Merleau-Ponty calls *maximum grasp* [23], which "varies with the goal of the agent and the resources of the situation" [4, p. 250]. Such a process underlies the pattern recognition that leads to human know-how and allows us to adapt to different situations in which we may be more or less certain of our knowledge about the world.

Because attempts at AI proceed from rationalist assumptions, computers have become adept at exactly the kind of abstract reasoning that Descartes believed to be characteristic of all knowledge. Thus, for example, computers can perform complex mathematical calculations much more rapidly and easily than can most human beings. They also succeed in situations such as games, which are able to provide clear reinforcement about whether the rules applied are correct or incorrect. The example of chess is relevant here as well. The triumph of computers over human chess grandmasters, as exhibited in the victories of IBM's Deep Blue over Garry Kasparov, has led some proponents of AI to believe that machine learning will surpass human intuition elsewhere. However, in chess the feedback from the world is binary: win or loss. Such clear reinforcement is rarely involved in the type of situated know-how at which humans characteristically excel. As Dreyfus argues, although some utilitarian moral philosophers have proposed to interpret all moral reasoning as a utility-maximizing mathematical function—the kind of calculation at which a computer would excel—human beings are rarely able to foresee all the consequences of a moral decision in a way that would allow for this sort of accounting [4, p. xlv]. Human practices in which conclusions are less certain seem to rely more upon the embodied, situated knowledge that Dreyfus describes.

In an attempt to overcome the limitations of the rationalist model of AI, some computer programmers have designed AI machines based on a *neural net*, a form of AI designed to approximate the way in which humans learn the kind of pattern recognition that underlies know-how. Instead of giving the computer pre-formed propositional knowledge, neural net programmers feed the computer raw data and allow it to make its own connections spontaneously, in a manner analogous to the way in which human beings seem to learn. Yet such programs have nevertheless encountered problems for the reasons that Dreyfus anticipates. One problem is that the rules guiding these spontaneous connections must still be written out propositionally for

the computer. On a deeper level, however, because AI machines necessarily lack a human body with human perception, they seem incapable of determining relevance in the same way that humans do. In other words, they make spontaneous connections between inputs that no human would ever make.

The results of Google's DeepDream AI project provide a vivid example of the differences between human and AI perception [24]. DeepDream is an image-analysis program based on a neural net. Its programmers intended for it to learn to determine what an image "looks like" by picking out patterns in images and using a form of probabilistic logic to reinforce them, in much the same way that a human might think that a cloud looks like a boat or a turtle. Yet when asked to process simple images, DeepDream found all sorts of random objects where none was apparent to the human eye. For example, when given a picture of President Obama, DeepDream produced an image with eyeballs, birds, and psychedelic whorls of color scattered across the president's face and the background. It found within the image groups of pixels resembling these objects and then simply interpreted them as such. These results, while rather dazzling to behold, show how different the computer's probabilistic grasp of similarity is from that of the human mind, and how difficult it is for a computer to "think" the way humans do without human faculties of perception. Lacking a human body, AI machines may be destined to produce knowledge that, to humans, proves more odd than helpful.

Since human beings cannot step outside their own pre-conceptual background knowledge so as to describe it propositionally, they cannot impart the features of this knowledge to a computer in terms of explicit rules. Without such rules, a computer will remain without the capacity to distinguish relevant features of the world like humans can. An AI computer may be able to identify connections between its inputs that humans cannot see, and some proponents of AI have argued that, in this way, AI can contribute to our knowledge about the world. More likely, however, is that, lacking a human pre-conceptual understanding of relevant knowledge, it would produce connections that are uninteresting at best and unintelligible at worst.

Dreyfus's analysis draws our attention to the importance of know-how, the type of human practical knowledge that underlies both our common-sense grasp of the world and our ability to acquire skills. Computer engineers have struggled to reproduce this type of knowledge in AI machines not only because it is difficult to represent propositionally, but also because it relies upon various forms of pre-conceptual background knowledge, such as sensory perception and socialization, that seem uniquely human. As shown below, humans thus remain particularly well suited for practices, such as primary care, in which know-how plays a central role.

Human know-how and primary care

Dreyfus's critique casts doubt on the notion that AI machines can replace primary care physicians, since machines lack access to the sources of background knowledge underlying the practice of medicine. Much of clinical knowledge consists of know-how, the sort of pattern recognition that Dreyfus describes. The structure of medical education in fact presumes the importance of experiential learning. The popularity

of problem-based learning in the preclinical years derives from a general recognition that bedside clinical reasoning is context-dependent, requiring a wealth of specific prior experiences to draw on. Physicians typically require abstract conceptual knowledge only secondarily, as a resource to consult when they encounter a situation that does not fit the pattern [25]. The purpose of the clinical years and residency training is to expose trainees to enough cases that they begin to perceive patterns in the diagnosis and treatment of disease. Such experience is especially crucial for primary care physicians, who often evaluate patients with new or nonspecific complaints. Dreyfus's analysis indicates that this type of clinical know-how cannot be adequately captured in propositional terms and will thus be difficult, if not impossible, to transmit to a computer.

One example of such know-how is the ability to take a "history of present illness" from a patient. As medical educators know well, medical students often fail to elicit clinically relevant information from patients despite spending more time than anyone else actually taking the history. By contrast, an experienced clinician can obtain a complete history in a fraction of the time. Such physicians would likely have difficulty explaining this ability discursively, for it rests on their capacity to perceive relevance. As the patient begins to tell his or her story, the physician immediately begins to think of other symptoms to ask about and physical exam maneuvers that might lead to a particular diagnosis. When I take a history, at no point do I resort to a rule such as "always consider asthma and ask about cough when a patient complaints of difficulty breathing." Rather, I recognize what parts of the patient's history are most significant, often by recalling similarities between this patient's complaints and those of many other cases I have seen. True medical expertise consists in precisely this type of experiential knowledge, not simply in the mastery of context-free abstract concepts of disease and treatment.

Dreyfus's critique also reinforces the importance of socialization and somatic experience for clinical know-how. Social skills, bodily experience, and imagination shape our pre-conceptual expectations and thus the way we perceive the world—which for primary care physicians consists of patients and their complaints. In order to apply its algorithms, an iDoctor would necessarily render this complex perceptual environment as knowing-that, thereby overlooking important information, such as the veracity of a patient's account. Any experienced clinician knows not to rely solely on patients' verbal reports of their illnesses. Some patients understate the burden of their symptoms, for example, and others lie outright about the purpose of their visit. Over time, a human physician might learn to attend to the subtle nonverbal cues, such as averted glance or speech disfluency, that can signal the withholding of information. In such cases, the physician might ask more questions to discover the truth or decide on a safer treatment plan. Yet an AI machine, without the resources of human experience or interpretive skills, must take the patient's history at face value, as an accurate transmission of propositional knowledge, potentially leading to mistakes in diagnosis and treatment.

Furthermore, even when patients are truthful and forthcoming, their narratives of symptoms are not just reports containing factual information; they are appeals to the physician as an embodied and social human being. They invite the physician to consult his or her own bodily experience and to imagine what such symptoms would

feel like and thus call forth a certain response. A physician can only truly understand the complaint of a patient with a headache if he or she has also suffered from headaches, or at least possesses the ability to draw on other experiences to imagine what such pain might be like. A patient describing a symptom or the impact of illness on his or her life might also make use of metaphor, relying on a set of shared social assumptions in order to convey meaning. Additionally, the appeal to a human physician invites the physician to respond with appropriate conversational expressions of empathy, such as facial gestures or consoling touches on the hand. These expressions, however minor, can often augment patients' healing response to treatment [26]. Even if an AI machine could perform these actions, it would likely have difficulty knowing how and when to use such gestural tokens appropriately, and still one wonders whether patients would have the same response to the touch of the iDoctor.

Just as importantly, a human physician can account for the patient's unique non-medical circumstances when devising a plan of care. If an elderly patient tells a physician that anti-hypertensive medication makes him so dizzy that he can no longer drive to visit his family, the physician can understand that information as relevant, having had a similar experience of the value of visiting with loved ones. Although the "best practice" guidelines for management of hypertension would recommend treatment, the physician might devise a more permissive blood pressure goal for this particular patient so as to better meet his needs. An AI machine, by contrast, does not possess similar memories of non-medical experiences to which patients might appeal. Dreyfus implies that an AI machine, lacking socialization within a human culture and the experience of embodiment, would necessarily fail to comprehend the nuances of these aspects of patient care.

Unlike games and other constrained situations at which computers excel, medicine rarely admits of enough certainty to allow for clear feedback. As Dreyfus points out, computers "learn" best when they are able to receive unambiguous reinforcement from a programmer or from the world with regard to the rules that they apply: right or wrong, win or loss. Yet outcomes in medicine are rarely so obvious or binary. Primary care physicians in office practice diagnose and treat patients without knowing for certain if they will return, and when they do return, their response to therapy is rarely a clear "better" or "worse." Many patients experience an ambiguous mix of benefits and side effects, leaving physician and patient to determine together the best way forward in a specific situation. Often the "right" answer, such as a statin medication for high cholesterol, is not the best answer for a particular patient's concerns. Primary care thus requires flexibility and adaptability, as well as a willingness to accept as sufficient the coping with the world that Merleau-Ponty calls maximum grasp [23]. Dreyfus points out that computers tend to have more difficulty applying this type of reasoning when they are faced with an increasing variety of circumstances, whereas humans tend to improve on their application. Any designer of an AI machine that purports to replace physician judgment would need to overcome this significant obstacle.

The ongoing attempt to produce driver-less cars provides some evidence of the problems that occur when AI technology is introduced into a human social milieu. A recent article aimed at the lay public notes that private and public investment in this type of AI has been massive, with GM alone spending $500 million and the

U.S. Transportation Secretary proposing to invest $4 billion [27]. Yet, to date, no version of this technology has been able to approximate human *social intelligence*, the ability to guess effortlessly and correctly what other people on the road will do. Whereas research shows that "even very young children have exquisitely tuned senses for the intentions and goals of other people"—as the "core of uniquely human intelligence"—AI cars frequently fail to anticipate the actions of human drivers [27]. Furthermore, although both AI machines and humans make mistakes on the road, the machines make mistakes that no human would: "They will mistake a garbage bag for a running pedestrian. They will mistake a cloud for a truck" [27]. In other words, AI machines have difficulty with the type of know-how that comes naturally to human beings. They cannot always identify the humanly relevant information from among their many inputs, especially in situations that call for knowledge gained by socialization, precisely the problems that Dreyfus predicts. A communal human practice such as driving, and by extension medical practice, may be possible only because human beings share the same background knowledge.

It is conceivable that an AI machine might eventually be capable of providing certain aspects of primary care, such as preventative services, more efficiently than humans do. No doubt some patients go to physicians solely to obtain such appropriate treatment as efficiently as possible. Yet many others have sought not only medical therapy but also compassionate care from their relationships with primary care physicians. As Dreyfus's analysis shows, the patient's complaint of chest pain appeals to a fellow human being capable of a similar experience. As the physician responds, the patient can perceive not only the physician's effort to diagnose and treat the problem but also the bodily signals, both verbal and nonverbal, of the physician's concern for the patient. Thus, the relationship between patient and physician opens the possibility for true compassion, understood etymologically as "suffering-with." This affective response then motivates the physician to use the tools at his or her disposal, including medical knowledge as well as empathic listening and presence, for the good of the patient.

An AI machine might be vulnerable to various malfunctions that would require "treatment" by an expert. However, because a machine cannot become incarnate, and therefore cannot be vulnerable in the same way as humans beings, it cannot possibly suffer with its patients. As MacIntyre has argued, this acknowledgment of the vulnerability of the human body gives rise to certain virtues that contribute to human flourishing [28]. These virtues, such as compassion, are essential for a doctor–patient relationship that not only provides efficient diagnosis and treatment but also assures the patient that the physician cares for him or her. Thus, close attention to the human body is crucial both for clinical knowledge and for the moral practice of medicine. Dreyfus's account suggests that AI may remain ever incapable of attuning to the human body in a way that comes naturally to other human beings.

Conclusion: Primary care as focal practice

As noted above, Heidegger and Borgmann's critique of technology shows how technology tends to distort focal things. It thus suggests an alternative model for primary

care as a focal practice centered on a focal thing: the care of the patient. Both parts of this focal thing, the care and the patient, are of ultimate concern. Within such a practice, care must be understood holistically, as not only diagnosis and treatment but also compassion and fellowship in the face of illness—benefits a human physician may be equipped to offer. Indeed, because the physician is human, there is some degree of contingency in the doctor–patient relationship, for its outcome is not determined algorithmically in advance. This contingency seems anachronistic, if not alarming, from within the technological enframing that tends toward standardization and control, but it also preserves an opportunity for grace, as doctor and patient may experience healing as a gift [29]. Similarly, it allows physicians the opportunity to develop their knowledge and skill, achieving excellence in a distinctively human way. This notion of primary care, then, challenges physicians to design and use technology such as EHRs to extend rather than attenuate their ability to care for patients. It also holds medical educators responsible for their duty to cultivate such intellectual and moral virtues in their trainees.

This model of primary care as a focal practice upholds the importance of the individual patient. Instead of imposing a technological frame of law-like generalizations and best practices on the patient, physicians in such a practice let patients be themselves, in all their particularity. This attempt to see patients anew, outside the technological enframing, fulfills the Heideggerian hope for medicine as an ontological opening that brings forth health from the individual patient's life story, rather than isolating those aspects of the patient's life that can be therapeutically manipulated in order to deliver health as a commodity. Such a doctor–patient relationship seems possible only between two human beings. As Dreyfus points out, our shared humanity allows us to recognize the most relevant parts of our experience and thus to arrive together at the truth. In other words, as human beings, physician and patient are capable of caring about the same things, and this mutual understanding makes possible an appropriate response [30]. In some circumstances, efficient diagnosis and treatment are enough. However, in others, particularly those of chronic and terminal illness, more is required. After all, disease and death remind us of our frailty and expose as a fiction the technological conceit that we do or can control nature. Thus, those of us trapped in the technological enframing are apt to perceive them not as parts of our natural human life story but as existential threats [31]. In the face of such suffering, the physician might offer not merely more technological artifice but rather wisdom and compassion derived from his or her own experience of being human.

I have argued in this paper that AI should not, and perhaps cannot, replace the physician's role in providing primary care. Yet it must be conceded that these arguments could be overcome by actual technological developments. The science of AI may well progress to the point that it achieves the capacity to replace the judgment and intuition of primary care physicians, at which point patients might welcome such a change. Heidegger was famously pessimistic on this point, declaring in an interview that "only a god can save us" from the encroachment of *Gestell* [32]. If technology can undermine even human communication and empathy, as demonstrated in the introduction, then it seems likely to exert greater influence over other human practices, including medicine. If nothing else, then,

perhaps this paper serves to show what goods are at stake if, or when, we walk into our primary care provider's office only to find the iDoctor there waiting.

References

1. Parikh, Ravi B., Meetali Kakad, and David W. Bates. 2016. Integrating predictive analytics into high-value care: The dawn of precision delivery. *Journal of the American Medical Association* 315: 651–652.
2. Weisberg, Jacob. 2016. We are hopelessly hooked. *New York Review of Books*. http://www.nybooks.com/articles/2016/02/25/we-are-hopelessly-hooked.
3. Turkle, Sherry. 2015. *Reclaiming conversation: The power of talk in a digital age*. New York: Penguin.
4. Dreyfus, Hubert L. 1992. *What computers still can't do: A critique of artificial reason*. Cambridge: MIT Press.
5. Heidegger, Martin. 1993. The question concerning technology. In *Basic writings*, rev. ed, ed. David Farrell Krell, 307–341. San Francisco: HarperCollins.
6. Polanyi, Michael. 1958. *Personal knowledge: Towards a post-critical philosophy*. Chicago: University of Chicago Press.
7. Borgmann, Albert. 1984. *Technology and the character of contemporary life: A philosophical inquiry*. Chicago: University of Chicago Press.
8. Heidegger, Martin. 1991. *Nietzsche*, *Vol. 1: The will to power as art; Vol. 2: The eternal recurrence of the same*. Trans. David Farrell Krell. San Francisco: HarperCollins.
9. Heidegger, Martin. 1991. *Nietzsche, Vol. 3: The will to power as knowledge and as metaphysics; Vol. 4: Nihilism.* Trans. David Farrell Krell. San Francisco: HarperCollins.
10. Thomson, Iain D. 2005. *Heidegger on ontotheology: Technology and the politics of education*. Cambridge: Cambridge University Press.
11. Feenberg, Andrew. 1999. *Questioning technology*. New York: Routledge.
12. Heidegger, Martin. 1993. On the essence of truth. Trans. John Sallis. In *Basic writings*, rev. ed, ed. David Farrell Krell, 111–138. San Francisco: HarperCollins.
13. Verbeek, Peter-Paul. 2011. *Moralizing technology: Understanding and designing the morality of things*. Chicago: University of Chicago Press.
14. Chi, Jeffrey, Maja Artandi, John Kugler, Errol Ozdalga, Poonam Hosamani, Elizabeth Koehler, Lars Osterberg, et al. 2016. The five-minute moment. *American Journal of Medicine* 129: 792–795.
15. Dexheimer, Judith W., Thomas R. Talbot, David L. Sanders, S. Trent Rosenbloom, and Dominik Aronsky. 2008. Prompting clinicians about preventive care measures: A systematic review of randomized controlled trials. *Journal of the American Medical Informatics Association* 15: 311–320.
16. Goitein, Lara. 2015. Training young doctors: The current crisis. *New York Review of Books*. http://www.nybooks.com/articles/2015/06/04/training-young-doctors-current-crisis.
17. Caplan, Arthur L. 2016. Physician burnout is a public health crisis, ethicist says. *Medscape*. http://www.medscape.com/viewarticle/859300.
18. MacIntyre, Alasdair. 1984. *After virtue: A study in moral theology*, 2nd ed. Notre Dame: University of Notre Dame Press.
19. Brenner, Hermann, Christian Stock, and Michael Hoffmeister. 2014. Effect of screening sigmoidoscopy and screening colonoscopy on colorectal cancer incidence and mortality: Systematic review and meta-analysis of randomised controlled trials and observational studies. *BMJ* 348: g2467.
20. Bourdieu, Pierre. 1977. The objective limits of objectivism. In *Outline of a theory of practice*, 1–71. Trans. Richard Nice. Cambridge: Cambridge University Press.
21. Searle, John R. 1979. Metaphor. In *Expression and meaning: Studies in the theory of speech acts*, 76–116. Cambridge: Cambridge University Press.
22. Lenat, Douglas B., and Edward A. Feigenbaum. 1991. On the thresholds of knowledge. *Artificial Intelligence* 47: 250–285.
23. Merleau-Ponty, Maurice. 1962. *Phenomenology of perception*. Trans. Colin Smith. London: Routledge and Kegan Paul.

24. Auerbach, David. 2015. Do androids dream of electric bananas? *Slate*. http://www.slate.com/articles/technology/bitwise/2015/07/google_deepdream_it_s_dazzling_creepy_and_tells_us_a_lot_about_the_future.html.
25. Norman, G.R. 1988. Problem-solving skills, solving problems and problem-based learning. *Medical Education* 22: 279–286.
26. Brody, Howard. 2012. On talking and touching in medicine. *Journal of Pain and Palliative Care Pharmacotherapy* 26: 165–166.
27. Anthony, Samuel English. 2016. The trollable self-driving car. *Slate*. http://www.slate.com/articles/technology/future_tense/2016/03/google_self_driving_cars_lack_a_human_s_intuition_for_what_other_drivers.html.
28. MacIntyre, Alasdair. 1999. *Dependent rational animals: Why human beings need the virtues*. Chicago: Open Court.
29. Lake, Christina Bieber. 2013. *Prophets of the posthuman: American fiction, biotechnology, and the ethics of personhood*. Notre Dame: University of Notre Dame Press.
30. Reich, Warren Thomas. 1996. A new era for bioethics: The search for meaning in moral experience. In *Religion and medical ethics: Looking back, looking forward*, ed. Allen Verhey, 96–115. Grand Rapids: Eerdmans.
31. Hauerwas, Stanley. 1990. *God, medicine, and suffering*. Grand Rapids: Eerdmans.
32. Heidegger, Martin. 1976. Only a god can save us: *Der Spiegel*'s interview with Martin Heidegger. Trans. Maria P. Alter and John D. Caputo. *Philosophy Today* 20: 267–28

Medicine, Health Care and Philosophy
https://doi.org/10.1007/s11019-020-09939-2

SCIENTIFIC CONTRIBUTION

Should we have a right to refuse diagnostics and treatment planning by artificial intelligence?

Iñigo de Miguel Beriain[1,2]

© Springer Nature B.V. 2020

Abstract
Should we be allowed to refuse any involvement of artificial intelligence (AI) technology in diagnosis and treatment planning? This is the relevant question posed by Ploug and Holm in a recent article in *Medicine, Health Care and Philosophy*. In this article, I adhere to their conclusions, but not necessarily to the rationale that supports them. First, I argue that the idea that we should recognize this right on the basis of a rational interest defence is not plausible, unless we are willing to judge each patient's ideology or religion. Instead, I consider that the right must be recognized by virtue of values such as social pluralism or individual autonomy. Second, I point out that the scope of such a right should be limited at least under three circumstances: (1) if it is against a physician's obligation to not cause unnecessary harm to a patient or to not provide futile treatment, (2) in cases where the costs of implementing this right are too high, or (3) if recognizing the right would deprive other patients of their own rights to adequate health care.

Keywords Artificial intelligence · Right to refuse treatment · Health care · Patients autonomy

Introduction

In July 2019, *Medicine, Health Care and Philosophy* published an extraordinarily interesting article. Ploug and Holm (2019) argued the need to protect the right to refuse diagnostics and treatment planning by artificial intelligence (AI). Nevertheless, the authors showed the possibility of distinguishing between a strong version of this right, which would allow the holder to refuse any involvement of AI technology in diagnosis and treatment planning, and a weak version, which would only allow recognition of the claim to physician involvement in the diagnostic and treatment planning process. The authors seemed to favour the strong version of the right, albeit with limitations, when patients' objections are 'based on rational concerns about the systemic effects of AI use'.

In this article, I adhere to their conclusions, but not necessarily to the rationale that supports them. Instead, I criticize some of the weaknesses I found in the authors' arguments and provide some alternative arguments that might serve better to support their proposals. To this purpose, I will start by stating that it is not necessary to introduce a discussion on the weak version of the right, but on its extension. As Ploug and Holm correctly state, the right as such has been clearly recognized by the current European Union (EU) legal framework, even though we have yet to define its boundaries.

Instead, I will focus on the idea that we must adopt the strong version of the right to refuse diagnostics and treatment planning by AI, but subject to severe restrictions. To that end, I will separate myself substantially from Ploug and Holm's argumentation. First, I will argue that the idea that we cannot root this right based on a rational interest defence. I will show that this is not plausible, unless we are willing to judge each patient's ideology or religion and this is against fundamental principles included both in the Charter of Fundamental Rights of the European Union (2012/C 326/02) and in the General Data Protection Regulation. Instead, I will argue that the proposed right must be connected with values such as social diversity or individual autonomy and responsibility. Afterward, I will point out that that the scope of such a right should be limited at least

✉ Iñigo de Miguel Beriain
inigo.demiguelb@ehu.eus

[1] Chair in Law and the Human Genome Research Group, Department of Public Law, University of the Basque Country, UPV/EHU, Barrio Sarriena S/N, Leioa, Bizkaia, Spain

[2] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

Published online: 20 January 2020

under three circumstances: (1) if it is against a physician's obligation to not cause unnecessary harm to a patient or to not provide futile treatment, (2) in cases where the costs of implementing this right are too high, or (3) if recognizing the right would deprive other patients of their rights to adequate health care.

The current EU legal framework: a weak version of the right to refuse diagnostics and treatment planning by AI in the GDPR

The EU legal framework on the application of AI to human health is resolved by Article 22(1) of Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation, thereinafter GDPR), which reads *'The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.'*

Obviously, the wording of this clause clearly indicates that it is not possible to use AI if there is no human element involved in the decision-making process. Therefore, we can definitely hold that the weak version of the right invoked by Ploug and Holm has already been recognised by EU law. However, they are perfectly correct when they point out that its scope is yet to be defined. Indeed, the article does not make explicit is the degree of intervention that must be considered necessary to conclude that the requirement is covered (Mitchell and Ploem 2018). Therefore, it is particularly important to highlight the statement made by the Article 29 Data Protection Working Party (2018), an advisory body comprising a representative from the data protection authority of each EU member state, which played a prominent role in terms of interpretation of the Regulation until it was replaced by the European Data Protection Board (EDPB) under the GDPR. In 2017, the Party clarified the scope of the prohibition by stating that:

> 'The controller cannot avoid the Article 22 provisions by fabricating human involvement. For example, if someone routinely applies automatically generated profiles to individuals without any actual influence on the result, this would still be a decision based solely on automated processing. To qualify as human intervention, the controller must ensure that any oversight of the decision is meaningful, rather than just a token gesture. It should be carried out by someone who has the authority and competence to change the decision. As part of the analysis, they should consider all the available input and output data.'

Therefore, the legal framework is clear in one part: there is no room for solely automated decision-making, at least in the EU zone (Dreyer and Shulz 2019). However, the degree of concrete involvement of physicians in the final decision and the tools implemented to guarantee that this does not become a mere formality remains unclear. It is hard to see at the moment how we could avoid that physicians routinely adopt the recommendations made by artificial intelligence due to defensive medicine considerations, for instance. The implementation of the weak version of the right and its concretisation includes a wide range of options and we have not faced this issue yet. Thus, there is an urgent need for discussion on this essential point and Ploug and Hold are perfectly right when they claim to be in favour of it, since social concerns must play a fundamental role in the decisions made. Hopefully, this will help us to resolve the concrete degree of human involvement that the weak version of the right involves and the best ways to guarantee it.

Nevertheless, we can at least conclude that at the moment, the weak version of the right to refuse diagnostics and treatment planning by AI, that is, the 'claim to physician involvement in the diagnostic and treatment planning process', has been endorsed by the EU regulation (Wachter et al. 2017). However, this does not at all mean that a strong version of the right, that is, the right 'to refuse any involvement of AI technology in diagnosis and treatment planning', is against the EU regulation. Indeed, in the next section I will argue that such a strong version of the right works well with some of the values that are widely accepted in the EU context, and thus there are some good reasons to support it.

The argument for the recognition of the strong version of the right

One of the parts I found most disturbing in the article by Ploug and Holm is that in the section entitled 'Rational concerns and dystopies' they defend the idea that the strong version of the right must be based on the patient's rational fears and concerns. In fact, the authors make a great effort to demonstrate that if a patient raises an objection to the use of AI for those purposes on the grounds of a possible undesirable societal effect, then we should respect the patient's claim and recognize their right to refuse diagnostics and treatment planning by AI. In my opinion, this is an unfortunate argumentation, as it concedes, in the negative sense that, if there is no rational explanation of the reasons for refusing AI, then the strong version of the right does not apply. This implies assuming the need to situate our focus on the reasonableness of

a request, which means questioning the rationality of an ideology or a faith, an attitude that violates Article 21 of the Charter of Fundamental Rights of the European Union (CFR). Indeed, this is not a typical course of action, of course. Take Jehovah's Witnesses case, for example. Do we really protect their right to refuse certain treatments on the basis of the rationality of their beliefs? In the answer to this question is the reason for my rejection of the theses of the authors of the article I now criticize (Petrini 2014).

Of course, I do not believe that the right to refuse diagnostics and treatment planning by artificial intelligence must be considered as a part of the general right to refuse treatment. I think that Ploug and Hold argue in a very convincing manner that both rights are different. Instead, I think that the principles that refrain us from judging the ideology or religious beliefs that support refusing a treatment should also apply to the right that the authors of the paper that I am commenting are describing. Furthermore, I consider that there are no good reasons to oblige patients to declare the reasons why they are opposing the use of AI in the decision process, provided that the conditions I mentioned in the introduction, and that I will explore in the following sections, apply. If this were the case, then it would only be the patient who would suffer the consequences of his or her negative to use AI tools. Therefore, I cannot see any strong reason to oblige him or her to reveal any kind of information about his or her ideology. Moreover, that would be contrary to the principle of data minimization, an essential ethical principle that has been incorporated into the European Union's General Data Protection Regulation (GDPR). This principle means that data processing should only use as much data as is required to successfully accomplish a given task. Provided that we can base the right refuse diagnostics and treatment planning by artificial intelligence on reasons other than the rationality of a belief (as I will hold immediately), I do not think that we have any reason to oblige patients to reveal these very sensitive personal data. Instead, if the conditions mentioned apply, then the reasons that guide the patients' decision would be totally irrelevant, since the right would not be applicable.

Rather, I believe that we must opt for the strong version of the right based on value pluralism and the patient's autonomy and responsibility. Value pluralism means that 'people's views diverge about a range of fundamental questions, political ethical and religious. This diversity appears to be inevitable and irresolvable. It is not possible to determine a single correct view or set of values (Turner 2004). As a consequence, negotiation, tolerance and compromise are necessary' (Wilkinson and Savulescu 2018). Indeed, this value has been embedded in the EU Chart of Fundamental Rights in its Preamble ('The Union contributes to the preservation and to the development of these common values while respecting the diversity of the cultures and traditions of the peoples of Europe') and plays a key role in EU societies at present.

My second argument relies on the concept of autonomy, a concept that is certainly mentioned in the paper by Ploug and Hold, but in a quite different sense. They consider autonomy on the basis of our acting as rational beings. In my view, autonomy refers here to the capacity of the patients to make their own decisions according to their principles and values. Indeed, I think that respect of patient autonomy is guaranteed under Article 8 of the European Convention on Human Rights of 4 November 1950, which protects the right to private and family life. Thus, it serves as an excellent root to the right to refuse AI intervention in health care. This is probably due to the fact that I think that autonomy must be understood not only as a right to refuse a treatment, but to make decisions on the whole treatment process, as autonomy is rooted in the importance of self-government and freedom to live according to one's goals (Varelius 2006; Hartzband and Groopman 2009).

Therefore, I believe that the strong version of the right we are considering is directly connected with basic values such as patient autonomy and value pluralism and therefore it must be fully accepted in the EU context. Indeed, the focus should be on the reasons we could oppose or at least request the restrictive use of a right that is directly linked to these fundamental principles and values. What could be the reasons for defining boundaries to the right to refuse diagnostics and treatment planning by AI in its strong version? In my opinion, there are two: the need to reconcile this right with respect to physicians' ethical concerns and the costs it might involve for health care systems. I will analyse both in the next sections.

The argument of physicians' right to make an informed decision

First, one might oppose the right we are considering by stating that physicians are meant to have a say in the diagnostic and treatment procedures used in the development of their work. However, if we do not adopt a paternalistic approach to medicine, I doubt that this statement involves a general right for physicians to make decisions without considering the patient's values and interests. For example, in the case of Jehovah's Witnesses, we concede to such patients the right to decide on how surgery should be performed, not only the right to decide whether or not they want to undergo it. Thus, it does not seem reasonable to consider that patients cannot decide on the diagnostic and treatment tools and the possibility of avoiding AI for these purposes. However, I think that this general right only applies if this does not yield as a consequence a violation of the physician's right to not act against globally recognized medical ethical principles, such

as non-maleficence or beneficence, for example (Macklin 2003). And this might certainly happen under some circumstances if we recognize a strong version of the right.

Indeed, there are cases when a physician might be unsure on whether a concrete treatment may be futile or even harmful to a patient. Imagine, for example, that a patient with cancer requests chemotherapy, but the oncologist does not know if this could be really effective in this concrete case. Under these circumstances, AI might be the only means of making a decision about it. If the patient exercises the strong version of the right, physicians would have to face a situation in which they might infringe their ethical duties: they might finally act without knowing if the intervention will not cause harm or death, not to mention the futile use of public resources, even though it would be possible to solve this dilemma using AI tools. In my opinion, cases such this show that the right to veto could undermine physicians' right to use the most accurate resources available to ensure that they are not disregarding the essential ethical principles in health care I have mentioned earlier. Indeed, physician refusal to provide futile or harmful care is supported by the ethical principle of non-maleficence, which seems particularly relevant in the situation described (Luce 1995).

Ploug and Holm might point out that this same happens in the case of Jehovah's Witnesses and yet physicians have a duty to proceed with the alternative treatment, but I think that both scenarios are not at all the same. In the case of Jehovah's Witnesses, physicians are forced to choose an alternative treatment that, in any case, would always work better than no treatment at all or death. In the case described, it might perfectly happen that the treatment would help the patient face their cancer, but it might also cause unnecessary pain to the patient. Thus, physicians forced to provide treatment without using AI would be aware that they could be causing an avoidable harm, but they would also know that if they were to not provide the treatment, they could flout the beneficence principle. The question of certitude is key in such situations, but this is precisely what the patient opting for the right would be stealing from physicians, and I think this is unfair to physicians. Moreover, I think this is a misunderstanding of the informed consent framework (Paris 2010).

Therefore, we must conclude that the right of patients to not use AI in decisions about their treatment cannot be extended to the point of forcing doctors to act against commonly accepted medical ethical principles. This could be expressed either by establishing this circumstance as a limit on the exercising of the right, or by accepting the right to conscientious objection from the health professional who is to provide the treatment. In my opinion, it seems more reasonable to adopt the first option, because if the principle of non-maleficence is a basic principle in medical ethics, we should not think that its respect implies the need to invoke the right to conscientious objection. Furthermore, as the principles at stake are universally accepted, would it make any sense to finally put into practice a treatment that could be futile or harmful only because the patient manages to find a doctor who does not mind carrying it out? In my opinion, such a physician would be flouting the principle of non-maleficence on the basis of the alibi provided by the principle of patient autonomy.

To sum up, I consider that respect of health care workers' principles and values is a strong enough reason to conclude that the right to refuse diagnostics and treatment planning by AI cannot be an unlimited right. Indeed, it seems reasonable to think that if AI can determine whether a treatment will serve, or instead cause harm or be futile, it is the obligation of a physician to make good use of it, due to the prevalence of the non-maleficence principle, which overrides the autonomy of the patient. I concede that Ploug and Holm might be correct with respect to the idea that in the future some physicians "may come to be biased toward the decisions made by the IA technology and less sensitive to the particular preference and interests of the individual". However, if this were the case they would not be practising medicine according to the goals and standards of their profession and, thus, these types of attitudes should never become a legitimate boundary to the right we are discussing now.

Health care system sustainability and the rights of other patients

The second factor when considering the limits of the right to refuse diagnostics and treatment planning by AI are those that derive from the costs that the recognition of this right could cause to the health care system. Ploug and Holm address this issue in a senseful way in their paper and I cannot but adhere to what they state, even though I would like to make some remarks to it. I do not share their belief that in some circumstances, "allowing some patients to refuse AI involvement (…) might lead to cost savings because patients who are strongly opposed to AI would avoid seeking health care until their conditions have progressed to a serious state". I dare say that, in those cases, the use of AI in a previous state of health could lead to a better diagnosis and a more effective treatment, rendering unnecessary the costly treatment that we usually have to administer to a serious health state. To the contrary, I suspect that recognizing the strong version of the right will probably reduce the savings that the implementation of AI in health care might bring. However, I do not think that this fact, even if confirmed by evidence, should play a definitive role in order to oppose the recognition of the right. It is very common, in fact, for the exercise of a right derived from patient autonomy to harm public health. This happens, for example, if we accept

that a patient rejects an optimal treatment, opting instead for another that will ultimately would lead to higher public health costs.

However, we usually accept this result based on the defence of principles such as the need to respect the plurality of values in non-uniform societies or the importance of respecting each person's life plans. Thus, for example, no patient is forced to undergo a kidney transplant even though the alternative (that is, long-term dialysis) is much more expensive for the system. Nor has compulsory vaccination been introduced against influenza, even though this substantially increases healthcare costs. Nor, of course, are patients penalized in general for not strictly following the recommended treatment, even though this may lead to relapses and higher costs. Moreover, there are strong reasons that support such policies (Howard 2008; Schmidt 2007). I therefore understand that the concept of increased public health costs should not serve to veto, in general, the strongest version of the law we are analysing.

However, I believe that there are exceptions to this general rule. If a treatment is particularly costly, for example, it should not be administered without first having recourse to the advice of the AI if the efficiency of the corresponding predictive algorithm had been demonstrated. This usually happens in health systems, which set specific indexes for decision-making on financing treatments (which as in the UK happens with the incremental cost-effectiveness ratio [ICER]) (Nikolentzos et al. 2008). I believe that this type of threshold could perfectly well be applied even in the case of recognising this right, setting objective limits to its exercise.

However, I believe that the main objection against the strong version of the right we are analysing comes from other types of situations. More specifically, it is necessary to consider cases in which the exercising of this right would cause obvious harm to third parties, who would be deprived of adequate care as a consequence. This might be better understood through an example. Imagine that eight people want to gain access to very expensive treatment. Furthermore, the statistics show that only half of the patients with that concrete condition respond to the treatment in a minimally reasonable way, following the prevailing criteria for allocating resources in that health system. Interestingly, there is an algorithm capable of precisely guessing which of these eight people will benefit from the treatment at a reasonable cost and which will not. However, four of them refuse to have the AI used in the analysis of their specific case.

Imagine now that the AI is used for the other four and that the algorithm determines that two of them are not treatable under the underlying cost conditions. This means that there are six people left who are likely to enter into the final selection of the four candidates. If we believe that the appeal to the right should not lead to any discrimination against those who exercise it, it would be logical to draw lots among the remaining six. Thus, fortune would decide impartially who will and will not be treated. However, statistically, this would imply that at least one person capable of healing would be excluded and one for whom treatment is futile would be treated.

In my view, however, this final distribution of resources would be absurd. The logical approach would be to administer treatment to the two people for whom the AI has made an encouraging prognosis and to circumvent the other two candidates among the four who want to exercise their right to not have these mechanisms used to decide on their treatment. The opposite would be to arrive at an inefficient and unfair result based on personal ideology. However, if this is the case, then it is clear that the right we are talking about has to be limited based on the costs for the health system, the need to optimise resouce allocation, but above all, on the right of third parties to access efficient treatment. It could, of course, be pointed out that the case I have put forward is exceptional and should not serve as a rule. I do not think that is true. It is a case that arises every time there is a drug shortage, and we must design a system for allocating scarce resources among patients who are likely to take advantage of it (or not). In my opinion, if AI were able to suggest an efficient form of allocation, we should not allow the right to refuse treatment planning by AI to deny scarce health resources to patients who are able to benefit from it.

Acknowledgements Iñigo de Miguel Beriain's work was supported by the Government of the Basque Country, Grant IT-1066-16 and the EU Commission, H2020 SWAFS Programme, PANELFIT Project, research Grant Number 788039.

References

Article 29 Data Protection Working Party. 2018. ARTICLE29 Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679 (wp251rev.01). https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053. Accessed 20 July 2019

Charter of Fundamental Rights of the European Union (CFR), 2012/C 326/02. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT. Accessed 20 July 2019

Council of Europe, European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14, 4 November 1950, ETS 5. https://www.refworld.org/docid/3ae6b3b04.html. Accessed 18 Jan 2020

Dreyer, S., and W. Schulz. 2019. The General Data Protection Regulation and Automated Decision-making: Will it deliver? Potentials and limitations in ensuring the rights and freedoms of individuals, groups and society as a whole. Working Paper. Bertelsmann Stiftung. https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/GDPR.pdf. Accessed 11 Aug 2019

Hartzband, P., and J. Groopman. 2009. Keeping the patient in the equation—Humanism and health care reform. *New England Journal of Medicine* 361: 554–555.

Howard, B.M. 2008. First, do not punish: Individual incentives in health policy, virtual mentor. *AMA Journal of Ethics* 10 (11): 719–723. https://doi.org/10.1001/virtualmentor.2008.10.11.conl1-0811.

Luce, J.M. 1995. Physicians do not have a responsibility to provide futile or unreasonable care if a patient or family insists. *Critical Care Medicine* 23: 760–766.

Macklin, R. 2003. Applying the four principles. *Journal of Medical Ethics* 29: 275–280.

Mitchell, C., and C. Ploem. 2018. Legal challenges for the implementation of advanced clinical digital decision support systems in Europe. *Journal of Clinical and Translational Research* 3 (Suppl 3): 424–430.

Nikolentzos, A., E. Nolte, and N. Mays. 2008. Paying for (expensive) drugs in the statutory system: An overview of experiences in 13 countries. London: London School of Hygiene & Tropical Medicine. https://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_ 089990. Accessed 20 Aug 2019.

Paris, J.J. 2010. Autonomy does not confer sovereignty on the patient: A commentary on the Golubchuk case. *American Journal of Bioethics* 10 (3): 54–56.

Petrini, C. 2014. Ethical and legal aspects of refusal of blood transfusions by Jehovah's Witnesses, with particular reference to Italy. *Blood Transfusion* 12 (Suppl 1): s395–s401. https://doi.org/10.2450/2013.0017-13.

Ploug, T., and S. Holm., 2019. The right to refuse diagnostics and treatment planning by artificial intelligence. Medicine, Health Care and Philosophy. https://doi.org/10.1007/s11019-019-09912-8. [Epub ahead of print]

Schmidt, H. 2007. Patients' charters and health responsibilities. *BMJ* 335 (7631): 1188.

Turner, L. 2004. Bioethics in pluralistic societies. *Medicine, Health Care and Philosophy* 7 (2): 201–208.

Varelius, J. 2006. The value of autonomy in medical ethics. *Medicine, Health Care and Philosophy* 9 (3): 377–788.

Wachter, S., B. Mittelstadt, and L. Floridi. 2017. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law* 7 (2): 76–99.

Wilkinson, D., and J. Savulescu. 2018. Cost-equivalence and pluralism in publicly-funded health-care systems. *Health Care Analysis* 26 (4): 287.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Medicine, Health Care and Philosophy
https://doi.org/10.1007/s11019-019-09912-8

SCIENTIFIC CONTRIBUTION

The right to refuse diagnostics and treatment planning by artificial intelligence

Thomas Ploug[1] · **Søren Holm**[2,3]

© Springer Nature B.V. 2019

Abstract
In an analysis of artificially intelligent systems for medical diagnostics and treatment planning we argue that patients should be able to exercise a right to withdraw from AI diagnostics and treatment planning for reasons related to (1) the physician's role in the patients' formation of and acting on personal preferences and values, (2) the bias and opacity problem of AI systems, and (3) rational concerns about the future societal effects of introducing AI systems in the health care sector.

Keywords Artificial intelligence · Bias and discrimination · Rational concern · Right to refuse · Health care

Introduction

The use of artificially intelligent (AI) systems for medical diagnostics and treatment planning is being heralded as one of the great avenues forward for modern medicine. In a number of countries AI systems such as IBM's Watson or Google's Deepmind are being tested in different health care settings.[1]

Enthusiasts for these developments claim that there are strong reasons for implementing such systems. By taking into account vast amounts of diverse research and personal health care data and by the ability to process data and reach decisions much more quickly than humans, AI systems will in the near future lead to more *precise* and *efficient* diagnostics and treatment planning than what may be achieved by physicians.[2] There is growing evidence that this may be true (Weng et al. 2017; Bedi et al. 2015; Rajpurkar et al. 2011; Esteva et al. 2017; Leung et al. 2016; Dilsizian and Siegel 2014). If the claims are true an increased precision and efficiency of diagnostics and treatment planning will not only lead to better health care outcomes through early and precise detection and intervention but also to more cost-efficient health care services. The exact scope and strength of the arguments and evidence is open to question. We take no position on this, but the analysis in this article accepts *for the sake of argument* that they are true.

In this article we consider whether patients should be able to exercise a right to withdraw from AI diagnostics and treatment planning in their individual case and have diagnostics and treatment planning performed by a physician.[3] We first explicate the different shapes such a right could have and its relation to the right "not to be subject to a decision based solely on automated processing" is guaranteed by article 22 of the European Union's General Data Protection Regulation (GDPR) We then analyse three clusters of arguments which in different ways justify that patients should be able to exercise this right for reasons related to (1) the physicians' role in the patient's formation of and acting on personal preferences and values, (2) the bias and opacity problem of AI systems, and (3) rational concerns about the future societal

✉ Thomas Ploug
ploug@hum.aau.dk

Søren Holm
soren.holm@manchester.ac.uk

[1] Department of Communication, Centre for Applied Ethics and Philosophy of Science, Aalborg University Copenhagen, A. C. Meyers Vænge 15, 2450 Copenhagen, SV, Denmark

[2] Centre for Social Ethics and Policy, School of Law, University of Manchester, Manchester M13 9PL, UK

[3] Faculty of Medicine, Center for Medical Ethics, University of Oslo, Oslo, Norway

[1] In this paper we use the term 'AI systems' to cover both systems based on 'symbolic AI' and systems based on machine learning techniques such as deep learning and neural networks.

[2] In this paper we use the term 'physician'. Diagnostic and treatment decisions are also made by many other types of health care professionals, but we are focusing on medical doctors because they are involved in many of these decisions.

[3] There is a parallel issue raised by AI controlled treatment, e.g. AI controlled surgical robots, but this is outside the scope of this paper.

Published online: 29 July 2019

effects of introducing AI systems in the health care sector. We argue that the last of these clusters of reasons provide the strongest support for a strong, general right to withdraw from AI involvement in diagnostics and treatment planning. Throughout the analysis we consider counterarguments, relating to costs, practicality and the potential negative effects in relation to the rights of other patients.

AI systems may be involved in generating the general evidence base for medical decisions, e.g. by conducting automated meta-analyses. This use of AI systems is outside the scope of this paper for two reasons. First, it does not engage all three types of reasons that are involved in the individual patient case. Second, it is distinct by being a practice aimed at generating generalisable knowledge and not preference-sensitive, individualised decisions.

A right to withdraw from AI diagnostics

What would a right to withdraw from AI diagnostics entail? It may be argued, that a right to withdraw from AI diagnostics and treatment planning is already encompassed in the generally acknowledged right to informed consent, i.e. that neither diagnostic interventions nor treatments can be performed without the informed consent of the patient. A patient therefore already has a right to refuse AI involvement. But the patient also has a right to health care provision, and this right is not completely waived or extinguished if the patient refuses a specific health care intervention. An example is the well-known case of Jehovah's Witness refusing blood transfusion on religious grounds. Jehovah's Witnesses are in most countries offered alternative treatments, such as blood sparing surgery. Another example is the case of patients suffering from severe 'needle phobia'. Their health care would be individually designed according to the specifics of their needle phobia and the interventions they need in order to minimise the need for blood sampling, injections, and infusions etc. The right discussed in this article to withdraw from AI diagnostics and treatment planning mirrors the right given to patients in these examples. It is not only a negative right to refuse a particular type of intervention, but also a positive right to insist on and be provided with an alternative type of intervention, i.e. in this case diagnostics and treatment planning performed by a physician. The right we argue for is therefore not coextensive with the right to informed consent.

There are many different types of possible physician involvement in diagnostics and treatment planning, and it has to be explicated and justified what kind of physician involvement the right we are discussing entails. At one end of the spectrum physician involvement may simply mean that physicians *are involved* in some way. A little further along the spectrum it may mean that physicians evaluate the quality of AI recommendations. And at the opposite end of the spectrum it would mean that that physicians *take care of all* diagnostics and treatment planning without any AI involvement. The right therefore comes in weak and strong versions.

In some cases, AI functionality is an integral part of a particular diagnostic device, e.g. ECG equipment which will provide an interpretation of the ECG trace. It is likely that more and more devices will have such functionality and that it will in practice become impossible to completely avoid AI involvement in the diagnostic process. Nevertheless, there are many steps between a particular piece of diagnostic information, a final diagnosis and the development of a personalised treatment plan, and it makes sense to ask whether there is a right to insist on the amount of AI involvement in this process being minimised. This would still be a fairly *strong* right.

In the European Union the right "not to be subject to a decision based solely on automated processing" is guaranteed by article 22 of the General Data Protection Regulation (GDPR) if the decision significantly affects a person (Art. 22 GDPR 2018). This prohibition applies to "decisions that affect someone's access to health services", and human involvement has to be meaningful and "it should be carried out by someone who has the authority and competence to change the decision. As part of the analysis, they should consider all the relevant data" (Article 29 Data Protection Working Party 2016, pp. 21–22) How strong a right to physician involvement this is depends on the interpretation of the requirement to consider all the relevant data. If it implies that a physician reads the entire patient record, this is fairly strong right. If it only implies that a physician briefly scans the input to and output from an AI system, it is a much weaker right. It is clear, however, that the GDPR right is weaker than a strong right to refusal of AI involvement since the strong versions of this right entail that patients may insist that the entire diagnostic and treatment planning process is carried out by a physician without AI involvement.

The patient's medical and non-medical preferences and interests

AI systems are, ex hypothesi, likely to be highly precise and effective in finding the right diagnosis and proposing the best treatment for a patient. The best treatment can be understood both as the most medically effective treatment and as the intervention that will be most cost-effective given the particularities of the patient.

Finding the right treatment cannot, however, be reduced simply to such an exercise of medical expertise. It is also an ethical exercise. In finding the right treatment the patient's preferences and interests of relevance for treatment decisions

must be considered. The patient's preferences and interests may differ from those of the health care system. The physician must try to ascertain these preferences and interests and take them into account in reaching a treatment decision. The key question therefore is, if it is possible for the AI systems to take these interests adequately into account in their treatment planning?

Adequate identification of patients' medical and non-medical preferences

Without communicating with the patient, one may suggest that AI systems could take into account the medical and non-medical preferences and interests of patients on the basis of data from health records of various kinds and data from non-health related databases. Deriving personal, health related preferences from social media data is an active area of AI research (Ghani et al. 2018; Jiang and Yang 2016). AI methods have, for instance, been developed in order to provide individualised information in online health communities based on the individual's prior engagement with and contribution to threads in the community (Jiang and Yang 2016). There are several reasons for being sceptical about the adequacy of this as a general strategy:

(1) Limited availability of data about patients' medical and non-medical interests.
(2) Lack of updated data about patients' medical and non-medical interests.
(3) Privacy protected data about non-medical interests.

Patients have medical preferences. They may e.g. prefer treatment by pills rather than injections. Patients also have non-medical preferences of relevance for treatment planning. They may e.g. hold family-life to be valuable, where this entails a medical preference for treatments that minimise hospitalisation.

None of these preferences may be stated in the patients' records or in any available health care database, and data that will allow for an adequate prediction of such preferences and interests may be very limited. Even if data about patients' medical and non-medical preferences are available in health care databases, it could be outdated and thus not reflect the patients' current preferences and interests at a given point in time. Patients' preferences are often unstable, ill-informed and inconsistent (Thaler 2000; Sunstein et al. 2002; Tversky and Kahneman 1986). It is difficult—if not impossible—to determine reliably the patients' current and true preferences and interests solely on the basis of historic data in patient records or other databases. Data in non-health related databases, e.g. data from social platforms or consumer data of various kinds, may provide insights into patients' non-medical preferences and interests. Access to such data is, however, protected by privacy regulation, and it is unclear whether AI systems in the health care sector will be allowed to access such data. Data from social platforms are also notoriously unreliable as indicative of real preferences since such platform 'performance' involves a large element of 'impression management' (Krämer and Winter 2008). Physicians can engage patients in discussion about their current preferences and help patients to understand how these relate to their treatment options. They therefore do not have to rely on previously collected data.

AI systems may, one may venture, at some point in the future be able to engage in communication with patients and credibly imitate the role of the physician as sketched above (Stein and Brooks 2017), and this will lead to more accurate predictions about the patients preferences and interests relevant for treatment planning. In order to achieve this, AI systems would have to be able to engage a patient in open-ended and meaningful conversation about all aspects of human life. Think, for instance, of a consultation with a woman who has been investigated for breast-cancer where she has to be told that she has cancer and have an initial discussion about treatment options. In such a consultation all aspects of human life—biological, psychological, social, sexual, spiritual etc.—may be relevant in order to understand what is important to her in relation to the very significant decisions she will have to make. A human doctor will be able to do this, but no currently existing AI system can. As long as AI systems do not have this ability to engage in open-ended, meaningful conversation with patients about their preferences, some patients may have a good reason to prefer the involvement of physicians.

The bias, discrimination and opacity problem of AI

The introduction of AI systems raises particular ethical concerns around bias and discrimination and the control of bias and discrimination.

The potential of biases in AI systems' decision-making has already been extensively discussed (Mittelstadt et al. 2016; Mittelstadt and Floridi 2015; Dilsizian and Siegel 2014; Bozdag 2013). While bias is not necessarily an ethical problem, it may lead to discrimination defined as ethically problematic bias that violates considerations of justice and equality (Friedman and Nissenbaum 1996; Schermer 2011; Bird et al. 2016; Calders and Verwer 2010). This can happen in at least three ways. First, bias may be inherent in the algorithms governing the AI systems in their diagnostics and treatment planning. Algorithms may thus favour overdiagnosis and overtreatment to underdiagnosis and under-treatment. Second, bias may ensue from systems imitating human decision-making simply because such decision-making

occasionally may be biased, e.g. by training systems on test data tagged by humans. Physicians may be prone to discrimination on the basis of age, and if the system is trained using a set of cases where age-bias occurs the system may end up with the same bias. Third, bias may unwittingly result from AI systems' decision-making. Imagine an AI system trying to make decisions that maximises health care outcome. The best possible treatment may require a high degree of patient compliance. There is evidence suggesting that patient compliance is associated with age such that patients in the age group 60–70 years old are the most compliant and the middle-aged among the least compliant (Dunbar-Jacob and Mortimer-Stephens 2001; Morrell et al. 1997). Thus, an AI aimed at maximising the health care outcome may use this evidence and end up not offering the best possible treatment to middle-aged patients and thus de facto discriminate against such groups.

The problem of bias in AI systems' decision-making is exacerbated by the opacity of many of these systems, i.e. that the decision-making procedure cannot be fully explicated (Burrell 2016; Zarsky 2013, 2016). We can detect possible bias in the pattern of decisions made by AI systems, but if we cannot understand the decision making algorithm it becomes much more difficult to prove that the system is biased in ways that will lead to ethically problematic discrimination. A statistical analysis may show that a system provides different treatment advice for young and old patients. This is possible bias and possible discrimination, but it may also reflect treatment relevant differences between age groups. If we cannot understand the decision-procedure of the system other than in purely mathematical terms, it becomes very difficult to distinguish between problematic bias and appropriate age-related treatment advice.

That AI systems may discriminate in ways that are difficult to detect constitutes a reason for granting individual patients a right to withdraw from AI diagnostics. It may be argued that there is trade-off between the benefits of using AI systems and potential discrimination in the form of unequal treatment. However, the state is committed to non-discrimination and equal treatment and can therefore not make such a trade-off. We could not imagine a minister of health publicly declaring that we have implemented an AI system with great benefits, but which is known to be discriminating against a particular group of citizens. A trade-off can only legitimately be made by individuals, who decide to "take their chance" with a potentially discriminatory AI system, and therefore individuals must be granted a right to decide whether to make this trade-off.

There is one further consideration in relation to the problem of the potential bias and discrimination of AI systems. This is that this problem is not specific to AI systems. Physicians may, as already noted above, be biased in their decisions too. Such biases may unintentionally lead to discrimination against particular groups of patients. However, there are many informal mechanisms that play a regulatory role in relation to physicians' decision-making. Physicians are part of a health care community. Often diagnostics and treatment planning are the outcome of the work of a health care team, and unless the biases are shared within the team they are likely to be corrected. Furthermore, any instance of intentional discrimination by a physician is likely to attract the attention of fellow physicians, and many hospitals/medical societies have policies for whistleblowing. Also, physicians must obtain informed consent from the patients before commencing treatment. The requirement of informed consent dictates that physicians shall provide the patient with adequate information about an intervention. Thus, the physician typically will have to face the patient and thereby allow the patient to engage in critical examination of the physician's decision-making, and potentially require a second opinion. Finally, most physicians have received training in medical ethics and health law. They are acquainted with the principles and legislation that apply to their profession, and in many countries they have even pledged to act justly.

There are also a number of formal mechanisms that play a regulatory role in relation to physicians' decision-making. Patients can complain about physicians' decision-making. Formal complaints may end up having a number of grave consequences for the physician including losing their job, being expelled from professional bodies and losing their license to practice medicine. Some jurisdictions require physicians to revalidate at regular intervals and this formal process often involves providing information from patients and members of their health care team on the behaviour of the physician in question. Although none of these informal and formal structural measures ensure non-discrimination by physicians, they confer significant costs on a choice to discriminate.

AI systems cannot at present meaningfully participate in the social relations that are the basis for the informal and formal mechanisms that regulate physician behaviour. But, the developers and manufacturers of the systems can. Developers and manufactures are also subject to informal mechanisms of control, e.g. in the form of potential reputational damage, if an AI system is found to make discriminatory decisions. Whether these mechanisms are as effective as the mechanisms regulating physician behaviour is open to question and ultimately must be settled empirically.

If patients have good reasons to believe that the mechanisms for detecting and avoiding bias and discrimination from AI systems are weaker than the mechanisms for controlling physician behaviour, then they must be able to refuse diagnostics and treatment planning by AI systems.

Rational concerns and dystopies

A third argument in favour of a right to refuse diagnostics and treatment planning by AI systems is based on the patients' fears and concerns. Patients may simply fear AI technology and therefore resist being diagnosed and have treatments suggested by such systems. This may take many forms. It may concern this technology as applied to their specific situation in the clinical setting, or it may concern the implications of the use of AI for the future state of society (Ipsos 2017). Among the possible undesirable societal effects are:

(1) That AI systems outmatch physicians which in turn reduces the level of human contact and care, or leads to the deskilling of physicians (Cabitza et al. 2017; Hoff 2011; Vellido 2019).
(2) That AI diagnostics and treatment planning become monopolised with a number of negative effects (Powles and Hodson 2017; Hayashi et al. 2018; Bostrom 2017).
(3) That AI systems take control of key institutions in society and become hostile towards humans (Bostrom 2014; Dave 2014, 2016; Müller and Bostrom 2016).

In the following we will focus on whether rational concerns about undesirable societal effects of the introduction of AI in health care can justify a right to withdraw from AI involvement in diagnostics and treatment planning.

Let us define rational concern the following way: A person X has rational concern about a future state of the world Y if and only if:

X believes that Y is undesirable,
X believes that Y may occur,
X can provide a coherent justification for how Y may result from the current state of society,
the occurrence of Y is supported/not ruled out by existing scientific evidence, and
the undesirability of Y is supported/not ruled out by a value-system which satisfies minimal requirements of public reason.

Rational concern is essentially a matter of being able to provide a consistent explanation of how the society may end up in an undesirable state that corresponds with scientific evidence and the reasonable judgement of a group of informed people (Habermas 2018; Rawls 2005). Conditions 3 and 4 entail that a societal development must not only be logically possible, but at least minimally plausible in light of current knowledge and scientific evidence. It is logically possible that a violent revolution will break out in Denmark tomorrow, but there is no coherent justification for how this may come about given what we know about the current state of Danish society and the available evidence concerning the preconditions of revolution. Condition 5 entails that a societal development must be considered undesirable by a group of people who are able to provide reasons in public discourse that are intelligible to others. A completely idiosyncratic evaluation of a state of society as undesirable would therefore not count as a rational concern. Note that so defined rational concern is directed at a state of society that may not be a direct threat to the individual—the individual may not even come to experience it. Note also that so defined (1) rational concern is a matter of degree, and (2) it may or may not cause emotional distress to the individual.

This account of rational concern implies that concern about how the introduction of AI systems may transform the health care system in the future is rational. It is rational to be concerned that AI systems may outmatch physicians, that AI diagnostics and treatment planning may become monopolised, and that AI systems may take control of key institutions in society. For all of these points of concern we can coherently explain how and why they may lead to an undesirable state of society, this explanation corresponds with scientific evidence and is endorsed by prominent researchers (Cabitza et al. 2017; Hoff 2011; Vellido 2019; Powles and Hodson 2017; Hayashi et al. 2018; Bostrom 2014, 2017; Dave 2014, 2016; Müller and Bostrom 2016), and the predicted outcome is consistent with a view held by a group of citizens of what is an undesirable state of society.

A right to act on rational concerns about a future state of society

There are several reasons for believing that individuals should be granted a right to act on rational concern about the systemic effects of introducing AI technology in the health care system, i.e. a right to insist on different degrees of human involvement in diagnostics and treatment planning. More specifically, there are at least five reasons why rational concern about the harmful, societal effects of introducing AI technology in the health care system should be accommodated, and these are:

(1) Democratic reasons
(2) Reasons of autonomy
(3) Reasons of solidarity
(4) Consequentialist reasons
(5) Epistemological reasons

In a democracy policy-making should be sensitive to the worries of citizens. Rational concern as defined above about the introduction of new AI technology in health care is not idiosyncratic, it is shared by groups in society and

amenable to public reason. The patients' rights and opportunities in a modern health care system are and should be sensitive to the views of the public in order to maintain its democratic legitimacy. This is not to say that public opinion should dictate the rights and opportunities in the health care system. By granting patients a right to act on their rational concern about potential harmful, societal effects of introducing AI systems in the health care sector, patients are given an opportunity to express publicly shared concern. Obviously, there are other ways such concern may be democratically recognised, e.g. through elections and voting. It is unlikely, however, that the introduction of AI in health care will ever become a major feature of the election platforms of political parties, which means that citizens cannot in reality express their rational concerns through their voting behaviour. A right to withdraw from AI diagnostics would grant patients an opportunity to express their rational concern. Also, a right to withdraw from AI diagnostics and treatment planning is a right that a patient can exercise without infringing the right of other patients to benefit from AI. Implementing AI without the right to opt out entails the refusal to recognise the rational concerns of some citizens as legitimate.

To act on rational concern is to exercise rational agency and autonomy. By granting patients a right to act on their rational concern about potential harmful, societal effects of introducing AI systems in the health care sector, they are not only provided with an opportunity to protect themselves against the distress that fears and concerns may generate. They are also provided with an opportunity to consider in their choices harmful, societal effects that are not recognised as such in the health care system. In short, such a right would promote not only autonomy as a right to self-protection against harm, but also rational agency and autonomy in the sense of acting rationally on one's conception of the good life and the good society.

In the health care system, we accommodate some patients' fears and concerns even when they are not shared by the majority and not rational by the definition given above. We have mentioned examples such as Jehovah's Witnesses and patients with needle phobia above. These accommodations reflect a solidarity-based case-by-case approach to patients' fears and concerns: Such fears and concerns may strike people in different degrees at various different times for various different reasons, but in general we maintain on grounds of solidarity a health care system that is capable of accommodating such fears and concerns through the offer of alternative health care provision. But, if we maintain a solidarity-based approach in cases of what many see as irrational fears and concerns, should we not also insist on a solidarity-based approach in cases involving rational concern?

There are also good consequentialist reasons for giving patients a right to act on their concerns about the harmful, societal effects of new AI technology. The potential harmful societal effects of introducing such technologies are manifold and ultimately may affect generations across space and time. It is not just worries that certain more or less narrow sectional interests will be violated sometime in the future. By giving patients a right to object AI involvement, the health care system introduces a mechanism that allows individuals to signal their concerns which may prevent these negative societal effects if enough people choose to act on their concerns.

Finally, rational concerns are epistemologically sensitive to evidence. The patient's rights and opportunities in the modern health care system are and should be based on whether or not there is evidence in favour of an option and its alternatives. There are strong ethical reasons for this. By giving patients a right to act on their rational concern about potential harmful, societal effects, they are given a right to act on concern that is sensitive to evidence.

A weaker or stronger right to withdraw from AI diagnostics in the future?

We have in this article explored three clusters of arguments to show that patients should be granted a right to withdraw from AI diagnostics and treatment planning, but have taken no position on which of the many possible versions on the spectrum from a weak to a strong right is best supported by the arguments.

The arguments concerning the physician's role in ascertaining the patients' preferences imply that a physician must be involved in the diagnostic process and the treatment planning as long as AI systems are unable to engage in meaningful conversations with the patients about their preferences. Similarly, the arguments concerning the problems of bias and discrimination in AI technology only sustains a claim to physician involvement in the diagnostic and treatment planning process. Taken together the two arguments support a right to demand that physicians are actively engaging with patients about their preferences, and that any output from an AI system is scrutinised by physicians prior to implementation.

The argument from rational concern points to a more extensive right to refuse any involvement of AI technology in diagnostics and treatment planning. This justification is not primarily based on a worry about AI involvement in "my" particular treatment, but a concern about the systemic effects of AI introduction and use in the health care system.

The arguments presented in this article therefore show that there is a right to withdraw from AI diagnostics and treatment planning, and that this is a strong right in the cases where it is based on rational concerns about the systemic effects of AI use. The health care system should therefore allow patients to act on this right and take it into account when implementing

AI systems. This will entail implementing AI systems in such a way that it is possible to opt-out of AI involvement in diagnosis and treatment planning. However, like most rights this is not an absolute right. It may be limited by other concerns. Let us briefly sketch two issues.

First, there are both financial and practical considerations that may provide reasons to limit the accommodations that must be offered to patients that exercise a strong right to withdraw from AI involvement. Allowing some patients to refuse AI involvement may in some circumstances confer costs on the health care system because alternative provision will have to be made. In other circumstances it may lead to cost savings because patients who are strongly opposed to AI would avoid seeking health care until their conditions have progressed to a serious state. The net cost, if any, of implementing a right to withdraw from AI diagnostics and treatment planning depends on the exact implementation and cannot be predicted. It is ultimately an empirical question that cannot be settled a priori. It is also important to note that even if it was established that implementing a right to AI withdrawal increases costs this would not be a conclusive reason to limit such a right. We would need to balance the importance of the right against the magnitude of the costs. In the long run we would also need to consider to what extent we would need to maintain legacy competences. Is there, for instance, a duty to educate the next generation of physicians how to treat patients without the help of AI? This is, as we have indicated above, a question for future research, but may perhaps be elucidated through consideration of a possible analogy with a duty to educate all doctors to be able to function without access to the diagnostic equipment that is only available in the richer parts of the world.

Second, there may be a risk that the distinction between a weaker and a stronger right will collapse in practice. Imagine that in the future AI technology will be more precise and efficient in diagnostics and treatment planning than most doctors. It will make decisions that are more likely to maximise the health benefits for the individual patient. In such a scenario, the literature on automation bias indicates that it is unlikely that physicians can remain wholly uninfluenced by the diagnostics and treatment planning of the AI technology. In making recommendations about treatment the physicians may come to be biased toward the suggestions made by the AI technology, and less sensitive to the particular preferences and interests of the individual (Goddard et al. 2012, 2014). If so, this can only be completely avoided if physicians do not have access to the information provided by the AI technology.

References

Art. 22 GDPR. 2018. Automated Individual Decision-Making, Including Profiling|General Data Protection Regulation (GDPR). https://gdpr-info.eu/art-22-gdpr/. Accessed 11 July 2018.

Article 29 Data Protection Working Party. 2018. ARTICLE29 Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679 (wp251rev.01). https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053. Accessed 20 June 2019

Bedi, G., F. Carrillo, G.A. Cecchi, D.F. Slezak, M. Sigman, N.B. Mota, et al. 2015. Automated Analysis of Free Speech Predicts Psychosis Onset in High-Risk Youths. *npj Schizophrenia* 1: 15030.

Bird, S., S. Barocas, K. Crawford, F. Diaz, H. Wallach. 2016. Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI. Rochester: Social Science Research Network. Report No.: ID 2846909. https://papers.ssrn.com/abstract=2846909. Accessed 29 May 2018.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*, 352. Oxford, New York: Oxford University Press.

Bostrom, N. 2017. Strategic Implications of Openness in AI Development. *Global Policy* 8 (2): 135–148.

Bozdag, E. 2013. Bias in Algorithmic Filtering and Personalization. *Ethics and Information Technology* 15 (3): 209–227.

Burrell, J. 2016. How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms. *Big Data & Society* 3 (1): 2053951715622512.

Cabitza, F., R. Rasoini, and G.F. Gensini. 2017. Unintended Consequences of Machine Learning in Medicine. *JAMA* 318 (6): 517–518.

Calders, T., and S. Verwer. 2010. Three Naive Bayes Approaches for Discrimination-Free Classification. *Data Mining and Knowledge Discovery* 21 (2): 277–292.

Dave, Lee R.C.-J. 2014. Hawking: AI Could End Human Race. BBC News. https://www.bbc.co.uk/news/technology-30290540. Accessed 15 Aug 2018.

Dave, Lee R.C.-J. 2016. Stephen Hawking—Will AI Kill or Save? BBC News. https://www.bbc.co.uk/news/technology-37713629. Accessed 15 Aug 2018.

Dilsizian, S.E., and E.L. Siegel. 2014. Artificial Intelligence in Medicine and Cardiac Imaging: Harnessing Big Data and Advanced Computing to Provide Personalized Medical Diagnosis and Treatment. *Current Cardiology Reports* 16 (1): 441.

Dunbar-Jacob, J., and M.K. Mortimer-Stephens. 2001. Treatment Adherence in Chronic Disease. *Journal of Clinical Epidemiology* 54 (12, Supplement 1): S57–S60.

Esteva, A., B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, et al. 2017. Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature* 542 (7639): 115–118.

Friedman, B., and H. Nissenbaum. 1996. Bias in Computer Systems. *ACM Transactions on Information Systems* 14 (3): 330–347.

Ghani, N.A., S. Hamid, I.A. Targio Hashem, E. Ahmed. 2018. Social Media Big Data Analytics: A Survey. Computers in Human Behavior. http://www.sciencedirect.com/science/article/pii/S0747563218304144X. Accessed 15 Feb 2019.

Goddard, K., A. Roudsari, and J.C. Wyatt. 2012. Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators. *Journal of the American Medical Informatics Association* 19 (1): 121–127.

Goddard, K., A. Roudsari, and J.C. Wyatt. 2014. Automation Bias: Empirical Results Assessing Influencing Factors. *International Journal of Medical Informatics* 83 (5): 368–375.

Habermas, J. 2018. *Inclusion of the Other: Studies in Political Theory*, 341. Hoboken: Wiley.

Hayashi, S., K. Wu, B. Tangsatapornpan. 2018. Competition Policy and the Development of Big Data and Artificial Intelligence. The Roles of Innovation in Competition Law Analysis. Accessed 15 Feb 2019. https://www.elgaronline.com/view/edcoll/9781788972437/9781788972437.00016.xml

Hoff, T. 2011. Deskilling and Adaptation Among Primary Care Physicians Using Two Work Innovations. *Health Care Management Review* 36 (4): 338–348.

Ipsos, M.O.R.I. 2017. Public Views of Machine Learning. Report Title:92.

Jiang, L., C.C. Yang. 2016. Personalized Recommendation in Online Health Communities with Heterogeneous Network Mining. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, 281–284.

Krämer, N.C., and S. Winter. 2008. Impression Management 2.0: The Relationship of Self-esteem, Extraversion, Self-efficacy, and Self-presentation Within Social Networking Sites. *Journal of Media Psychology: Theories, Methods, and Applications* 20 (3): 106–116.

Leung, M.K.K., A. Delong, B. Alipanahi, and B.J. Frey. 2016. Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets. *Proceedings of the IEEE* 104 (1): 176–197.

Mittelstadt, B.D., P. Allo, M. Taddeo, S. Wachter, and L. Floridi. 2016. The Ethics of Algorithms: Mapping the Debate. *Big Data & Society*. https://doi.org/10.1177/2053951716679679.

Mittelstadt, B.D., and L. Floridi. 2015. The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. *Science and Engineering Ethics* 22 (2): 303–341.

Morrell, R.W., D.C. Park, D.P. Kidder, and M. Martin. 1997. Adherence to Antihypertensive Medications Across the Life Span. *Gerontologist* 37 (5): 609–619.

Müller, V.C., N. Bostrom. 2016. Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In *Fundamental Issues of Artificial Intelligence*. Cham: Springer, 555–572. (Synthese Library). https://doi.org/10.1007/978-3-319-26485-1_33.

Powles, J., and H. Hodson. 2017. Google DeepMind and Healthcare in an Age of Algorithms. *Health and Technology* 7 (4): 351–367.

Rajpurkar, P., J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, et al. 2017. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-rays with Deep Learning. [cs, stat]. http://arxiv.org/abs/1711.05225.

Rawls, J. 2005. *Political Liberalism*, 589. New York: Columbia University Press.

Schermer, B.W. 2011. The Limits of Privacy in Automated Profiling and Data Mining. *Computer Law & Security Review* 27 (1): 45–52.

Stein, N., and K. Brooks. 2017. A Fully Automated Conversational Artificial Intelligence for Weight Loss: Longitudinal Observational Study Among Overweight and Obese Adults. *JMIR Diabetes* 2 (2): e28.

Sunstein, C.R., D. Kahneman, D. Schkade, and I. Ritov. 2002. Predictably Incoherent Judgments. *Stanford Law Review* 54 (6): 1153–1215.

Thaler, R.H. 2000. From Homo Economicus to Homo Sapiens. *The Journal of Economic Perspectives* 14 (1): 133–141.

Tversky, A., and D. Kahneman. 1986. Rational Choice and the Framing of Decisions. *The Journal of Business* 59 (4): S251–S278.

Vellido, A. 2019. Societal Issues Concerning the Application of Artificial Intelligence in Medicine. *KDD* 5 (1): 11–17.

Weng, S.F., J. Reps, J. Kai, J.M. Garibaldi, and N. Qureshi. 2017. Can Machine-Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data? *PLoS ONE* 12 (4): e0174944.

Zarsky, T.Z. 2013. Transparent Predictions. *University of Illinois Law Review* 2013: 1503.

Zarsky, T. 2016. The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making. *Science, Technology and Human Values* 41 (1): 118–132.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Medicine, Health Care and Philosophy
https://doi.org/10.1007/s11019-020-09948-1

SCIENTIFIC CONTRIBUTION

A critical perspective on guidelines for responsible and trustworthy artificial intelligence

Banu Buruk[1] · Perihan Elif Ekmekci[1] · Berna Arda[2]

© Springer Nature B.V. 2020

Abstract
Artificial intelligence (AI) is among the fastest developing areas of advanced technology in medicine. The most important qualia of AI which makes it different from other advanced technology products is its ability to improve its original program and decision-making algorithms via deep learning abilities. This difference is the reason that AI technology stands out from the ethical issues of other advanced technology artifacts. The ethical issues of AI technology vary from privacy and confidentiality of personal data to ethical status and value of AI entities in a wide spectrum, depending on their capability of deep learning and scope of the domains in which they operate. Developing ethical norms and guidelines for planning, development, production, and usage of AI technology has become an important issue to overcome these problems. In this respect three outstanding documents have been produced:
1. The Montréal Declaration for Responsible Development of Artificial Intelligence
2. Ethics Guidelines for Trustworthy AI
3. Asilomar Artificial Intelligence Principles
In this study, these three documents will be analyzed with respect to the ethical principles and values they involve, their perspectives for approaching ethical issues, and their prospects for ethical reasoning when one or more of these values and principles are in conflict. Then, the sufficiency of these guidelines for addressing current or prospective ethical issues emerging from the existence of AI technology in medicine will be evaluated. The discussion will be pursued in terms of the ambiguity of interlocutors and efficiency for working out ethical dilemmas occurring in practical life.

Keywords Artificial intelligence (AI) · Ethics guidelines · Ethical values · Technology ethics

Introduction

At first glance, artificial intelligence (AI) reminds us of artificial machines perceived as merely sci-fi which might take over humanity one day. However, whether we are using its products consciously or not, AI can be found in everyday life. It is in our smartphones providing touch/face recognition and other guiding assistance, for example, anticipating users' next actions based on the current action by referencing their habits; in our smart cars with self-parking features; in navigation systems suggesting efficient routes to destinations we search for by using a neural network system; even a mobile robot navigating an unknown environment (Patle 2018). AI is inevitably used for social media to send personalized notifications to users' timelines by factoring in their past web searches and interactions or other types of users' mobility on that social media platform. It is effective in automated customer support applications which help us find a particular product we want to buy, and in the finance sector such as detecting credit card fraud, measuring credit risk, and robo-advisory (Wall 2018). AI is also used in the education field to customize educational content, create innovative teaching methods, and facilitate communication between students and lecturers (Chassignol et al 2018). Last but not least, healthcare systems derive benefits by using AI

✉ Banu Buruk
banuburuk@gmail.com

Perihan Elif Ekmekci
drpelifek@gmail.com

Berna Arda
berna.arda@medicine.ankara.edu.tr

[1] Department of History of Medicine and Ethics, TOBB ETU University Medical School, Ankara, Turkey

[2] Department of History of Medicine and Ethics, Ankara University Medical School, Ankara, Turkey

Published online: 31 March 2020

technology for digital consultations and proper medication management for patients (Jiang et al. 2017). AI enables physicians, healthcare providers, and pharmaceutical experts to achieve better results in the health sciences such as advanced diagnosis, personalized medicine, and drug design. In the fields of surgery and radiology, especially for complex surgery interventions, AI helps to capture and process large amounts of image data. In the field of cardiology, AI is helpful for cardiac imaging by segmentation and identification of health issues, the classification of images cataloged by different sources, and lesion detection (Dorado-Diaz et al. 2019). For hospital administration processes, medical records are kept digitally by efficient and accurate AI applications, resulting in the ability to provide real-time patient statistics information both to the physician and the patient (Haleem et al. 2019).

Although AI is being used in many different fields of human life, its definition is interchangeable as no general theory of intelligence or learning exists which would unite the discipline (Moor 2006). Before the 1950s, computers could be instructed to perform tasks but were unable to store information about their actions (Anyoha 2017). Alan Turing, who was a mathematician and logician often referred to as the father of modern computing, asked the crucial question, "Can machines think?" in 1950, which opened the path to assigning human intelligence features to machines, in other words the start of AI technology improvement (Turing 1950). The thinking process is dependent on cognition and intelligence features, and these are characteristic of human beings and AI which is a non-living entity.

The concept of AI was first introduced into the literature in 1956 at an historic conference titled "Dartmouth Summer Research Project on Artificial Intelligence" which was the event that initiated AI as a research discipline (Moor 2006). Since then, the progress of AI technology has been accelerated by improvement of machine learning algorithms. AI technology should be able to simulate the features of "human brain intelligence," characterized as learning, predicting, analyzing, and creating solutions (Say 2018). As a general definition, AI technology is the process of developing systems which are endowed with the intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalize, or learn from past experience abilities that can be assigned to a machine. Consequently, research in AI has focused mainly on the following components of intelligence: learning, reasoning, problem-solving, perception, and using language (Artificial Intelligence 2019). According to the Future of Life Institute, the AI that can be found currently is the form of narrow/weak AI which is described as "the technology designed to perform narrow tasks such as only facial recognition or only internet searches or only driving a car" (Future of Life 2016). Researchers are working on the improvement of today's AI capabilities to create a strong, or general AI (AGI) which has all the cognitive abilities that are performed by a real human brain (Future of Life 2016). According to the Montreal Declaration for a Responsible Development of Artificial Intelligence,

> *AI technology has the possibility to create autonomous systems capable of performing complex tasks of which natural intelligence alone was thought capable: processing large quantities of information, calculating, and predicting, learning and adapting responses to changing situations, and recognizing, and classifying objects." A European Commission document titled A definition of AI: Main Capabilities and Scientific Disciplines defines AI as "the system[s] that display intelligent behavior by analyzing their environment and taking actions – with some degree of autonomy – to achieve specific goals* (MDRDAI 2018b).

According to the above-mentioned application areas and definitions, AI enables users to obtain safer, more accurate, and effective results in a shorter time and make accurate inferences about the future. So, the advantages of AI are more about time, security, accuracy, and foresight; all of which create a comfort zone to today's people. Lifestyles and speed of technology are dependent on this comfort area which is provided by AI by self-managing many choices and decision-making processes, as a result of gaining a form of personhood situation. Such dependency and personal conditions place AI in a more effective and powerful position, and thus ethical standards are required. To this end, developing ethics guidelines for the planning, development, production, and usage of AI technology is crucial. In this respect there are three outstanding ethics guideline documents for AI as follows:

1. *The Montréal Declaration for Responsible Development of Artificial Intelligence* released by the Montréal University with the participation of non-governmental organizations (NGOs), academicians, specialists, and policy developers in 2018. (Hereinafter referred to as MDRDAI.)
2. *Ethics Guidelines for Trustworthy AI* published by the European Commission in 2019. (Hereinafter referred to as EGTAI.)
3. *Asilomar Artificial Intelligence Principles* were developed and published in conjunction with the 2017 Asilomar Conference. (Hereinafter referred to as AAIP.)

AI, as all the other emerging technologies; could adapt very quickly to our daily lives, we now use it more and more, and therefore we face more ethical problems. However, there are also new norms and values brought upon by AI; its` ability to improve its original program and decision-making algorithms via deep learning abilities. This autonomy makes

AI unique among all other advanced technologies in terms of ethical issues. Confidentiality, security, responsibility, equality, accountability and transparency are at the forefront of ethical issues associated with AI. In this sense, different stakeholders and people from different disciplines started to voice AI related subjects. The first document, which was released by Montréal University where a meaningful amount of AI research is being conducted, aims to orient the development of AI to support the common good, and guide social changes by making recommendations in a collective manner. The way to the preparation of the Montréal Declaration for a Responsible Development of Artificial Intelligence, firstly in November 2017 The Forum on the Socially Responsible Development of Artificial Intelligence was held in Montreal. This forum was arranged to invite the public to contribute in discussing the issues of responsible artificial intelligence. In the following years, from 2017 to 2018 people having different backgrounds such as computer science, ethics, and other related disciplines participated in meetings, discussions, and philosophy workshops The Quebec Commission on Ethics in Science and Technology had also participated to these discussions. As a result in December 2018, the Montréal Declaration for a Responsible Development of Artificial Intelligence by Montréal University has been prepared on the basis of the discussions held by different people from expertise fields (MDRDAI 2018a).

The second document, published by the European Commission and first presented by the High-Level Expert Group on AI, defines "trustworthy AI" in three dimensions: lawful, ethical, and robust. According to the European Commission's "united in diversity" motto; AI has a great potential to cause socio-economic changes; so there is a need to get prepared for these changes by the brainstorming of more than 500 contributors. Ethics Guidelines for Trustworthy AI has been prepared by the contribution from public and the High-Level Expert Group on Artificial Intelligence, and it is continuously open for revisions." (EC 2019).

The third document, which was developed during the 2017 Asilomar Conference, is a representation of awareness of the changes resulting from AI technology which affect every person in society, underlying the shared responsibility of all people in and out of AI research. The preparation process of The Asilomar Principles of Future of Life Institute had been started with the sense among the 2017 Conference on the Future of Artificial Intelligence attendees. Due to rising awareness of AI, many important reports have emerged from academia (e.g. the Stanford 100-year report), government (e.g. two major reports from the White House), and the nonprofit sector. The opinions in these reports have been gathered, checked for any overlaps and scored to create the 23 principles (Future of Life 2017a).

Recently various agencies have made efforts to produce guidelines to regulate the ethics of AI. Out of these, we selected the above three because of their unique features, namely, that they were written in collaboration by professionals from various backgrounds who have a wide area of influence, regulative powers, and who are highly regarded by academicians and professionals alike.

In this study, we analyzed these three documents based on the ethical principles and values they contain. Subsequently, we analyzed each document's approach to ethical issues and ethical analysis of the values to which they refer when one or more of these values are in conflict. We also analyzed these documents according to their citations to the bioethics field, their content acquisition, their impact, and future scenarios they draw.

Methods

All three AI ethics guidelines examined in this article express different values at different dimensions from research to consumer use, and from human rights to environmental sensitivity. Our justification for the selection of these AI guideline documents is mainly based on the decision on selecting policy documents including guidelines those are expressed in a normative ethical concept. We searched Google, Google Scholar and Web of Science databases for the guidelines themselves and articles on "artificial intelligence ethics", "artificial intelligence principles" and "artificial intelligence guidelines". We shaped our literature search according to the inclusion/exclusion criteria as follows:

1. We excluded the documents those were written more than 3 years ago, all the three documents are current and up to date.
2. We chose to make a comparison between issuers from different environments. Thus, the issuers of the selected documents reflects different point of views, "Montréal Declaration for a Responsible Development of Artificial Intelligence" reflects the academic background, "The Asilomar Principles" of Future of Life Institute reflects the combination of private sector backgrounds and European Commission's "Ethics Guidelines for Trustworthy AI" document reflects the governmental and political background.
3. We excluded the national goals of countries and the AI documents prepared by national governments; we rather preferred to study international sights.
4. The selected documents are all written in the countries where AI technology is being developed and the AI research is funded with meaningful budgets. In the literature, it is possible to encounter different AI guidelines from different countries such as Japan, China and Australia. However, according to our selection criteria 3; we

excluded the guidelines that mainly refer to a national context.
5. The authors of all the selected documents are from different backgrounds, we excluded the evaluation of the documents reflecting the limited number of expertise.
6. The selected documents are the widest range of documents in terms of criteria based on ethical values they touch upon.
7. The selected documents are the documents are not written and covered, but are documents that are open for revisions over time.
8. All documents selected are original documents being frequently cited sources in the literature. The guideline documents selected do not qualify for the continuation of other documents, and do not reference the documents written before than them.

After the document selection, we analyzed their content with a matrix prepared according to their subtitles. We found that a few similar topics are included in different titles of each document. In addition, we identified the themes of each document that do not correspond to other documents.

The selection of the criteria given in Table 1 for comparison is based on the brainstorming process with the above methodological process. We focused on criteria that are compatible with each other and refer to the documents. Then we studied the issues on which all three documents are focused, such as equality, prioritization of justice, and non-discrimination in the development and use of AI technologies. Following this, we searched for the possibility of revision of the documents as AI technology evolves. We also studied the ways in which AI is portrayed by the documents, whether it is perceived as a tool or as a purpose, and also whether or not they have a human-centric approach. Finally, we analyzed the terminology created for the development and implementation of AI technology and its equivalent in the documents and end with conclusion remarks.

Table 1 Criterion used for the comparison of guidelines

1. Date	8. Security
2. Authors	9. Utility
3. Target group	10. Equality
4. Referenced ethics values	11. Bioethics citation
5. Human control and responsibility sharing	12. Future scenarios
6. Autonomy on the basis of personal data	13. Ethics analysis methods
7. Transparency and globalization	

Results

All the documents have been written within the same timeframe and are also current. Therefore, the development process of AI technology applies to all three documents. Although researchers from different fields of specialization took part in the writing stages, opinions of people with different statuses were taken and opinions were exchanged among many stakeholders. However, as most computer science experts are concentrated in certain areas, that is, more developed countries, some geographic areas such as Africa, South and Central America, and Central Asia do not have equal participation in the preparation of these guidelines, thus developed countries are more influential in shaping possible ethical dimensions of AI technology than others (Jobin et al. 2019). Still, both MDRDAI and AAIP are open to online signatures for anyone wishing to support these guidelines and, so far, 1583 researchers from AI and robotics and 3447 from other fields have signed the AAIP (Future of Life 2017b).

Criteria date, authors, and target groups

Table 2 below outlines the general properties of the guidelines, that is, the first three criteria we used for comparison: date, authors, and target group.

It was found that the issues on which all three documents are focused are similar to each other. All underline the need for equality, prioritization of justice, and non-discrimination in the development and use of AI technologies by preventing social inequalities, power imbalances, and lifestyle impositions. All the documents perceive AI as a tool, not as a purpose, but in different ways. MDRDAI declares AI as a tool for the well-being of humans which is stated in the first principle as, "The development and the use of AI systems must permit the growth of the well-being of all sentient beings" (MDRDAI 2018b). The second document, EGTAI, published by the European Commission, declares AI as a tool in the same direct manner as MDRDAI; pointing out that, "AI use should be in the service of humanity and the common good, with the goal of improving human welfare and freedom" (EC, 2019). The third document, AAIP, also perceives AI as a tool, albeit in two stages. The first principle of the Asilomar document declares that, "The goal of AI research should be to create not undirected intelligence, but beneficial intelligence" (Future of Life 2017b). The latter's first principle shows us that AAIP perceives AI as a tool for the good of mankind, but "beneficial intelligence" is reached in the second stage. The first stage is to create a superintelligence by avoiding the assumptions regarding upper limits on future AI

Table 2 The general properties of documents with regard to the first comparison criteria

	MDRDAI	EGTAI	AAIP
Publication Date	December 4, 2018	April 8, 2019	January 8, 2017
Authors	Researchers working in the fields of computer science, robotics, law, philosophy, ethics, public health, criminology, neurophysiology, and communication at the University of Montréal as a result of negotiations with public representatives, experts, and other stakeholders	A group of 52 senior researchers in the fields of AI, robotics, machine learning, technology ethics, law (academia, civil society, and industry)	The Asilomar Conference had a total of 146 academicians and participants from leading companies in the development of AI technology such as Tesla, Google, and DeepMind
Target Group	Anyone wishing to take part in the responsible development of AI, regardless of whether the contribution is scientific or technological	Anyone who wants to create an ethical framework for AI firstly in Europe and then in a global environment	Anyone interested in the subject, including R&D researchers in the development of AI technologies

capabilities. This means that the first purpose is to create a human-level AI intelligence, while the second purpose is to develop this superintelligence into a form which will share common ethical ideals for the common good. In this sense, the Asilomar AI principles do not undervalue the possibility of human-level AI. Furthermore, it perceives AI as a tool for the benefit of all humanity rather than one group of people or profit motivated organizations. In respect of the AI perception in the Asilomar document, the probability of future discrimination and non-equality is foreseen as a long-term issue. Accordingly, the AI perception of all documents shows that all three are in favor of human-centric AI. The human-centric feature of AI is directly related with the user-centric design for especially disabled people. Within the scope of the principles of equality and justice, it should be indispensable to design the AI technology developed to improve the well-being of all people including healthy and disabled persons. In fact, in this context, the potential power of AI should be supported in order to eliminate inequalities between people and to equip disabled people to access the benefits of technology. However, none of the documents touch upon the issues about AI usage by disabled people. In addition, all the documents are active documents, open to revision according to the accumulation of scientific knowledge and techniques in line with feedback from different actors such as users, scientists, lawyers, ethicists, etc.

Criteria referenced values

Our fourth comparison criterion is "the referenced ethics values" in documents. MDRDAI includes 10 principles; EGTAI includes 11 principles in a combination of four ethical principles with seven key requirements; whereas AAIP covers 23 principles. Table 3 below lists the principles of the three documents, respectively. In addition, Table 4 shows the principles of the three documents with the classification of the criteria between criteria 5 and 10; these are listed in Table 1 and are compatible with each other in all three documents. As a result, we classified the compatible principles in six categories. The first category we developed is "human control" which covers responsibility, human agency-oversight, responsibility, and human control. The second category is "autonomy on the basis of personal data" which covers respect for autonomy and protection of privacy-intimacy, respect for autonomy, prevention of harm and privacy-data governance, personal privacy, and liberty-privacy. The third category is "transparency and globalization" that covers democratic participation, explicability, transparency and accountability, and failure of transparency. The fourth category is "security" which covers prudence, technical robustness-safety, safety, and risks. The fifth is "utility" which covers well-being, societal and

Table 3 The principles included in the documents, respectively, MDRDAI, EGTAI, and AAIP

MDRDAI	EGTAI	AAIP
Well-being	Respect for human autonomy	Research goal
Respect for autonomy	Prevention of harm	Research funding
Protection of privacy and intimacy	Fairness	Science-policy link
Solidarity	Explicability	Research culture
Democratic participation	Human agency and oversight	Race avoidance
Equity	Technical robustness and safety	Safety
Diversity inclusion	Privacy and data governance	Failure transparency
Prudence	Transparency	Judicial transparency
Responsibility	Diversity, non-discrimination, and fairness	Responsibility
Sustainable development	Societal and environmental well-being	Value alignment
	Accountability	Human values
		Personal privacy
		Liberty and privacy
		Shared benefit
		Shared prosperity
		Human control
		Non-subversion
		AI arms race
		Capability caution
		Importance
		Risks
		Recursive self-improvement
		Common good

Table 4 Categorization of the principles included by the documents, respectively, MDRDAI, EGTAI, and AAIP (according to their compatibility)

	MDRDAI	EGTAI	AAIP
Human Control	1. Responsibility	1. Human agency-oversight	1. Responsibility 2. Human control
Autonomy on the Basis of Personal Data	1. Autonomy 2. Protection of privacy-intimacy	1. Respect for autonomy 2. Prevention of harm 3. Privacy-data governance	4. Personal privacy 5. Liberty-privacy
Transparency and Globalization	1. Democratic participation	1. Explicability 2. Transparency 3. Accountability	1. Failure transparency
Security	1. Prudence	1. Technical robustness-safety	1. Safety 2. Risks
Utility	1. Well-being	1. Societal and environmental well-being	1. Shared benefit 2. Shared prosperity 3. Common good
Equality and Justice	1. Solidarity 2. Equity Diversity inclusion	1. Fairness 2. Diversity, non-discrimination, and fairness	1. Judicial transparency 2. Non-subversion

environmental well-being, shared benefit, shared prosperity, and the common good. The final, sixth category is "equality and justice," which covers solidarity, equity, diversity inclusion, fairness, diversity, non-discrimination and fairness, judicial transparency, and non-subversion. There are also some principles unique to documents such as sustainable development for MDRDAI; and research goals, research funding, science-policy link, research culture, race avoidance, value alignment, human values, AI-arms race, capability caution, importance, and recursive self-improvement for

AAIP. EGTAI does not include any ethical principle that is not mentioned in the other two documents.

The categorization is mainly related to the criteria which are compatible with each other and refer to the documents. Although the documents use mostly the same terminology and the criteria in Table 4 refer to categories such as human control or transparency, each document deals with these categories from different perspectives. Thus, our criteria from 5 to 10 constitutes the categorization of principles under which we drew comparisons of compatible values.

Criterion human control

All three documents speak the same language in terms of human control during the use of AI technology; however, they all differ in their approaches. The differences in approach are more about the level of human control in action.

MDRDAI's approach is inclined toward reinforced human control over AI technologies. The document states that the decrease in human responsibility in decision-making processes with the development of AI technology should not be directly proportional. Even if the speed of technology development accelerates, AI human control should not be reduced. So MDRDAI underlines the notion that the final decision after an AI process should be taken by a "free and informed person."

EGTAI prioritizes human control in a different way. It states that human control and responsibility issues are important for the protection of autonomy. The first key requirement mentioned in the document, which is "Human agency and oversight," summarizes the need for human control over AI. According to this requirement, human control can be achieved through several processes. First, human-in-the loop (HITL) means the presence of human intervention in every decision cycle of the AI system. In fact, a continuous human intervention is not a feasible method and furthermore, it is against the logic of AI, the purpose of which is to make human work easier. As a result, HITL is not the preferred process. Second, human-on-the loop (HOTL) implies the capability for human intervention during AI system design and monitoring. The third possible process is the human-in-command (HIC) process. This refers to the human process of overseeing the activity of the AI system and human decision-making about when and how to use the AI system in any particular situation (EC 2019). The human-in-command process refers to human control at the level of the decision regarding the question, "For what purpose will the AI system be used?" According to the HIC process, the job of deciding at which stages of our lives and for what purpose AI will work should be left to the human himself. All HITL, HOTL and HIC processes envisage human control at a certain level of AI activity. According to EGTAI, there should be less human control over an AI system and there has to be stricter governance, which actually means a different kind of human control that is not in the AI system cycle but is rather an external one.

However, AAIP considers the potential for realization of decision-making cycles in which humans are not involved. AAIP moves one step further than the HIC process, where there is no human control over the purpose for which AI technology should be used. AAIP considers the possibility, even the reality, that AI intelligence will reach the level of human intelligence at some point in the future. Therefore, the "human control" principle in the document states that "Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives" (Future of Life 2017b). In short, AAIP states that before moving on to the age of self-controllable AI technology, people should work on designs that make it effective in compliance with ethical standards.

Criterion autonomy on the basis of personal data

In terms of respect for autonomy and the rights regarding personal data, all three documents approach the issue from different angles. First, MDRDAI says that everyone should protect their right to personal data. This means that deciding the fate of personal data should rest with the person who owns that data. EGTAI produces discourse on the same path as MDRDAI but adds that the personal data of vulnerable groups require special attention. On the other hand, AAIP states that people need to have control over the data they produce (not own) and that AI technology applied to personal data should not restrict people's freedom. However, AAIP does not express an opinion about human autonomy of his/her own personal data. It is certain that the main challenge regarding the use of personal data is the aggregation of data, in terms of defining the use of it that could directly affect the individual (IEEE 2019). The guidelines do not provide direct explanations for the aggregation of personal data used by AI systems, nor do they offer solutions for the orientation of such big data.

Criterion transparency and globalization

All three documents discuss transparency of different subjects. MDRDAI argues the transparency issue under the principle "Democratic participation." In essence, this principle underlines the importance of the transparency of the code of AI technology algorithms in front of public authorities. Here, the principle of transparency is being correlated with the principle of justice, i.e. saying that an equality between people from different backgrounds/countries/societies should be ensured. Still, the document says that some of the code of decision-making algorithms should remain

confidential; applying the principle of "as open as possible, as closed as necessary."

EGTAI mostly focuses on the issue of transparency and globalization under the principles of explicability, transparency, and accountability. According to the document there is a need for traceability, controllability, and reporting of the adverse impacts of AI technology for accountability/correction, and the protection of whistleblowers involved in these reports. The principle of explicability describes technical transparency; which means the necessity for understandability of the AI system by human beings. In other words, this means there should be no black box case related to AI applications. In this sense, the explicability principle is related to the democratic participation of MDRDAI which prioritizes the transparency of the code of AI technology algorithms. At this point, HITL, or HOTL, may be useful for such technical transparency. Moreover, as the document itself states, the second principle of transparency is closely related to the principle of explicability. The transparency principle mostly underlines the quintessence of traceability of AI systems while they are gathering, processing, and reasoning the data. The document notes that transparency and traceability are must-haves for the auditability and explicability of AI technology. Communication is another issue mentioned in the transparency principle of EGTAI; it refers to the requirement that AI systems should not look like real people. Users should be able to identify whether they are interacting with human intelligence or artificial intelligence. This level of transparency is important for the protection of autonomy because this is directly related to the right of a person to know the source of information presented in the decision process, or who or what kind of technological device manages it.

The third document, AAIP, deals with the transparency issue under the principle "failure transparency" by explaining the need to share unsuccessful results of AI technology groups with competitors. AAIP declares that high-tech companies should cooperate in their R&D processes during the commercial development of AI technology by sharing information as this would be ethically appropriate rather than storing, or hiding, information from each other during R&D.

Criterion security

MDRDAI analyzes the security issue under the prudence principle. As the title of the principle suggests, security should be ensured by foreseeing undesirable results. The principle itself states that "Every person in AI development must exercise caution by anticipating, as far as possible, the adverse consequences of AI use and by taking the appropriate measures to avoid them." (MDRDAI 2018b). The document both underlines the necessity for the projection of negative results and also the necessity for the identification of cases that are used outside the purpose of AI. Here the document also touches indirectly on the issue of malevolence.

EGTAI focuses on the subject of security under the principle "technical robustness and safety." According to the document, safety for both the AI system as well as those using it should be ensured. An AI system can only be safe if technical robustness is enabled. An AI system should be robust enough to protect the data of the user from operational or system interacting agents. On the other hand, EGTAI discusses "the level of safety" as a matter of fact. Level of safety means the "the level of accuracy"; in other words, the security level of the service offered by an AI technology. The document states that, "The level of safety measures required depends on the magnitude of the risk posed by an AI system, which in turn depends on the system's capabilities." (EC 2019). Achieving technical robustness is the result of concentrated R&D process of AI technology. The level of technical robustness achieved as a result of each R&D process may differ; some may be statistically more accurate, while some may produce less accurate results. Thus, the issue of safety is directly related to the trustworthiness of AI technology in this sense.

AAIP deals with the security issue by two principles; safety and risks. With these two principles, the document takes issue with future scenarios. The safety principle talks the same language with the prudence principle of MDRDAI, which states the need for taking appropriate precautions to avoid undesired results. Therefore, AAIP declares that AI systems should be inspected throughout their operational lifetime. In particular, AAIP mentions plans for measures that are directly proportional to risk, in order to mitigate catastrophe and existential risks.

Criterion utility

Although all three documents make different naming for the principles about utility, their contents are similar to each other. The well-being principle of MDRDAI mentions "growth of the well-being of all sentient beings." Thus, it would be true to say that the common good approach covers not just human rights but animal rights which should be regarded as a humanistic attend.

EGTAI's societal and environmental well-being principles also give credit to sentient beings as well the environment under the issue of sustainability. The document makes a correlation between the two principles of fairness and prevention of harm under the well-being principle. EGTAI also considers the well-being of future generations.

AAIP deals with the utility issue with three principles: shared benefit, shared prosperity, and common good. The shared benefit and shared prosperity principles declare a globalized approach, that is, benefit to all people or as many people as possible worldwide. While, on the other

hand, the common good principle addresses the possibility of superintelligence—a scenario which can be named as "AI beyond human intelligence." The AAIP's common good principle states that superintelligence should be developed for the benefit of all humanity, rather than a group of people who develop or own that technology.

Criterion equality and justice

MDRDAI deals with equality and justice issues with three principles: solidarity, equity, and diversity inclusion. In common, these principles talk about respect and tolerance to different lifestyles, cultures, and thoughts, as well as future generations. According to the Montréal document diversity is important for maintaining justice for the development and usage of AI technologies.

EGTAI considers the issue by addressing objectivity with two principles: fairness and diversity-nondiscrimination-fairness. First, the document explains the necessity of the balance between competing interests and objectives in terms of ensuring there is no unfair bias. The fairness principle also correlates this balance issue with the prevention of discrimination and stigmatization. The second principle mentions universal design, ensuring equal access through inclusive design processes in which the AI systems are user-centric. Here it states that an AI system should be accessible to all people regardless of their backgrounds, culture, age, gender, or restrictions. EGTAI promotes equality and justice by promoting accessible AI technologies for both children and adults, and healthy and disabled people, considering also the vulnerable group's right to benefit from the advantages of AI products.

AAIP analyzes equality and justice issues with two principles: judicial transparency and non-subversion. Judicial transparency means openness of the judicial decision-making process, and if this process is operated by an artificial intelligence then it should be monitored by a competent human authority. On the other hand, the principle non-subversion places responsibility on highly advanced AI systems to respect and improve the social and civic processes on which the health of society depends.

Criterion bioethics citation

The solidarity principle of MDRDAI cites bioethical issues in stating that AI systems that provide health care should be considered important in patient-physician relations. The document is still onside with humanistic technology, and heeds real human to human interactions.

EGTAI provides direct examples of the use of AI in the health care field to protect human life. Examples include assisting caregivers, supplying elderly care, monitoring patients' conditions, early diagnosis, and efficient drug design.

On the other hand, AAIP does not refer directly to bioethics, but the use of personal data is addressed under the principle of liberty and privacy stating that "The application of AI to personal data must not unreasonably curtail people's real or perceived liberty." This means that the personal health data of persons should stay confidential and should not change the status of the liberty of the person to whom that data belongs. Actually, some unique statements could be added to these guidelines, for example, the Belmont Report on human subject research, in terms of developing AI assisted or oriented health technologies (Goering and Yusta 2016). Furthermore, research developing AI technologies for improving human health should also have user-centered design in terms of helping researchers to note different perspectives of those with disabilities, from different cultures, or even from different generations (Guan 2019).

Today, AI technology is being used to analize large-scale information of DNA, proteins, and especially the protein–protein interactions. Such big data is being used for drug design, for accurate clinical decisions, for diagnostic tools, for developing preventive treatment methods and also for personalized medicine approach. Some of these AI implementation examples have been mentioned in EGTAI document's "Health and Well-being" title, however privacy and confidentiality issues have not been associated with the big data issue which is indispensable to AI research in healthcare. AI technology is data dependent; the more data, the faster AI can learn. AI can produce more accurate results this way. However, the data used to feed AI's learning process in medicine is health data, and this data is called as sensitive data, which can lead to unwanted results such as discrimination. Confidentiality prevents personal health data from being propagated improperly and misuse. Today, personal health data are now kept only in the locked locker of the physician who communicates with the patient, but in electronic environments where other medical staff such as technical people, hospital manager and molecular biologist can also reach. The technological conveniences provided by AI cause patient data to be more accessible and therefore more sensitive.

Another issue about AI usage in medicine is the long-term risks and benefits. In the documents, long-term benefit-risk analysis of AI implementation in health field has not been addressed. Actually this problem is a common problem for all emerging technologies such as the nano-toxicity of nano materials in nano-medicine (Madannejada et al. 2019). However, AI has a leading role among other emerging technologies by having self-learning capacity. This feature of AI allows it to be capable of determining potential risks and benefits for human health.

Moreover, none of the documents discuss the possibility of the substantial change in the concept of self, or human enhancement by AI technology scenarios. In the context of medical ethics, the standards of normal functioning, disease, ability expectation, treatment and enhancement should be highlighted. Another ethical issue about human enhancement by AI is the decision about who will have access to these opportunities, in terms of the principles fairness and justice.

Criteria future scenarios and ethics analysis methods

We could not have imagined 100 years ago that we would reach our current technological level, and today we may have difficulty predicting the maximum capacity of AI a further 100 years from now. Therefore, we should not assume that AI has an upper limit, and ethical principles should not be established according to this possibility. Believing that there may be an upper limit for AI means that we would have to limit its potential strength and potential benefits. In this context, the first two of the guidelines, MDRDAI and EGTAI, both apply the HOTL approach toward future AI; in these guidelines there is no room for superintelligence. On the other hand, the faster and more robust the connectivity of an AI system, the more autonomous it will become, in other words, it will move closer to a state of superintelligence (Chimuka 2019). AAIP argues for the possibility of a superintelligence one day in the future, and does not support placing an upper limit on AI technology.

Moreover, when we encounter an ethical dilemma about the possible consequences of an AI system, the guidelines may be helpful in identifying the possible options, but not necessarily helpful for evaluating and choosing the best option among them. Therefore, in terms of the ethical analysis methods of the guidelines there is no hierarchy among the ethical principles of these guidelines. For example, EGTAI states that at times the principle of prevention of harm and the principle of human autonomy may be in conflict. In such a situation, EGTAI advises that an evidence-based solution rather than intuition or a random one should be preferred. This stills leaves the principle evaluation process to AI developers and users.

Discussion

Today AI is weak in that it is specialized in only one specific task. General AI is certainly on the way, and future generations, or even we, will witness AI technology that can perform complex tasks in the same way as a real human brain. As a result, we believe that the potentiality of superintelligence must be considered. Such foresight is very important for the ethical dimensions of AI technology as discussed above, especially human control, transparency, and safety issues (Tegmark 2017; Bostrom 2014). Even though it is widely believed that human-level AI is not yet on the doorstep and the probability of creating an AI system capable of fully mimicking very complex/open-ended real human world remains in the future (Marcus and Davis 2019); we maintain that even a low probability rate should be taken into account. Such an approach is necessary in terms of ethics if the purpose is to set the ideals even for the worst scenario of future AI in order to prevent possible bad consequences for human dignity before they materialize.

Human controlled AI is inevitable for the MDRDAI and EGTAI guidelines, whereas AAIP looks beyond the human condition. Beyond human AI refers to the cognitive and moral sufficiency of an AI system which is able to conduct free decision-making processes. However, even AAIP does not support an AI system free of human control in the case of judicial decision-making processes. The issue of justice of AI technology draws a critical boundary around the decision about what is lawful or not, or the decision about who is guilty or not. Although new AI systems are being developed that can undertake some of the functions of the legal profession, the ethics guidelines still do not consider that AI technology will be among HIC processes (Simpson 2016).

Transparency is another important concluding issue. It should be ensured for AI systems at different levels such as protecting social and cultural diversity, standardizing ideas, sharing negative results, and being honest by not allowing AI itself to resemble a real person. Transparency is important for ensuring social and cultural diversity in some fields because otherwise there may be outcomes which cause some unwanted results. For example, the conditions for maintaining health data in countries may differ, which means that various sources of the same kind of data may have different levels of accuracy. An AI system using such data with different accuracy levels is likely to produce inaccurate results which could be disastrous in the health profession. Moreover, as the data grows, accuracy level differences could lead to more serious false results, which is why diversity inclusion should be ensured only if transparency is ensured. Transparency is also an important issue in terms of trustworthiness. It must be clear to end users that an AI system is not a real intelligence. Users have the right to know whether they are interacting with a real or an artificial entity, especially while sharing their personal data. A transparent AI system increases users' trust of AI systems (Lee et al. 2019). Thus, how and when data is collected by an AI system and how the personal data collected by the AI system will be stored should be open to users (Luxton et al. 2016). The relationship between transparency and understandability of AI systems is as critical as the relationship between respect for human autonomy and informed consent for human research

(Felzman et al. 2019). Therefore, this level of transparency is crucial for the protection of autonomy of a person and reliability of the technology.

Security has to be mentioned. The level of safety or the level of accuracy of an AI system is especially important. The issue of security is directly related to the necessity to not ignore differences in the accuracy level of data sources in order to preserve the diversity mentioned above under the transparency issue. A scoring system, such as 95% or 70% accurate AI, could be applied to address how the source data gathered and processed differ from each other in terms of projecting real life data.

Furthermore, some of the ethics terminology used for AI are considered according to the guidelines. The first terminology, "ethics *for* design," refers to the adaptation of ethical principles for researchers who are designing, using, and supervising AI technology. The second terminology, "ethics *in* design," refers to strategies developed for process management/design in the implementation phase where AI technology interacts with social parameters. The third terminology, "ethics *by* design," implies the integration of self-value-based decision-making capabilities into intelligent machines with AI technology (Dignum 2018). The first two terminologies, ethics for design and ethics in design, both highlight the idea that responsible AI requires a responsible person, that is, AI which is being developed within the scope of ethical values must be human controlled; consequently, these terminologies are associated with MDR-DAI and EGTAI. On the other hand, the third terminology, ethics by design, implies that AI is a teammate which shares responsibility with people, hence this terminology can be associated with AAIP.

Finally, it is necessary to address two additional subjects regarding the guidelines: content acquisition and balancing, and the enforcement power. The first of the final two important subjects is content acquisition and balancing, as well as specification. In general, the necessity of establishing a reflective equilibrium between ethical principles is not mentioned much and the assumptions about ethical dilemmas are not emphasized. It is not clear in any of the documents what content the generically written principles will be adapted to in specific cases and which principle will be spent in favor of the other in the case of a dilemma. Creating AI systems with a certain level of autonomy enables them to be in a position authorized to solve ethical dilemmas. In other words, an AI system, especially one dealing with personal data, must be familiar with solving ethical dilemmas, for example, where to sacrifice the principle of respect for autonomy for the principle to not harm; or any other possible ethical principle conflicts. Making these decisions is as important as writing the principles. Moreover, the specification issue is important in terms of the contribution of different fields of expertise to writing the documents. It can be seen that people with technical expertise, such as computer sciences and robotics, have more say in the documents. Since these documents are ethical guidelines for the development and use of AI technology, the voices of experts in technology ethics, philosophy, and sociology should be raised during their preparation.

Last but not least, the second important subject to be addressed is the enforcement power of the documents. Techniques such as machine learning used in the development of AI technology are not framed from a value/principle-based ethical perspective but rather, developed within the framework of economic logic. Since fast results are the most important criteria to the business world which seeks profit, ethical evaluations are often ignored, and documents remain at the level of wishes, thus weakening the enforcement power of the ethics guidelines of AI. Actually, enforcement power would be ensured with the help of legal arrangements that prioritize these ethics guideline documents. For example, in all human studies, whether clinical or not, in order to eliminate suspicion of possible ethical breaches in the research, approval of an independent ethics review committee is mandatory according to the legal regulations. These regulations are based on the international ethical regulations for research involving human subjects or any kind of human subject data (CE Oviedo Convention, 1997; WMA Declaration of Helsinki, 1964). Similarly, in order to eliminate a possible breach in ethics of AI products, legal regulations underpinned by these three AI ethics documents would make ethical approval before entering the market mandatory. Actually, the fact that these documents are written from an economic benefit perspective rather than an ethical value perspective, and the perception that content acquisition and balancing will be made in favor of economic benefits instead of values, is a serious ethical problem in and of itself. A solution for such possible ethical problems arising from the documents could be through the revision process of the documents with the help of feedback. Certainly these documents are kept worthy by being alive and open for revision.

Conclusion

AI technology has found a solid place in our lives and there is no doubt that it will keep its place. Since this technology is seated at the core of daily life by the strong influence of its decision-making algorithms, ethical issues begin to gain in importance. This has resulted in several ethics guidelines being prepared in order to inform people about the ethical aspects of AI technology. In our analysis of the three main AI ethics guidelines, we encountered similar language with slightly different content relating to some of the principles. According to the categorization of the principles included in the documents these are human control, autonomy, transparency, security, utility, and equality; all three documents

approached these aspects from a different point of view. Different levels of human control and different aspects of human autonomy, transparency, security, and equality issues are mentioned in the documents. The result of these different perspectives is that the documents have quite different future scenarios. What is common to all three documents are the issues of content acquisition and balancing, and enforcement of power. Since these documents are ethics guidelines for AI technology, there are no grounded suggestions for ethical dilemmas occurring in practical life; neither is a strategy for reflective equilibrium between ethical principles included in the documents. On the other hand, three of the guidelines are considered to be uninterested in addressing current or prospective ethical issues emerging from the existence of AI technology in medicine. Another important issue to be touched upon is the ambiguity of interlocutors for these documents. They are more likely to be informative documents to raise public awareness of AI and to guide AI users to be more conscious while using AI technology. Thus, the guiding statements of the documents for AI technology developers are more at an advisory level. Nonetheless, factors that could increase the impact envisaged in these documents are directly related with the cumulative degree of knowledge gained from AI R&D processes. The reality is that AI technology is data dependent, and to be able to transform into AGI would require an enormous amount of data. Gathering and processing this data will take a meaningful period of time and the perception of AI itself, as well as its ethical challenges and guidelines, will evolve during this time period.

*This article is based on an oral presentation of the authors titled "A Critical Perspective on Guidelines for Responsible and Trustworthy Artificial Intelligence" and presented at the X. Turkey Bioethics Symposium: Advanced Technologies in Health and Ethics, October 17–18, 2019 in Istanbul.

References

Anyoha, R. *Can Machines Think?* 2017. https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/. Accessed 24 November 2019.

Artificial Intelligence. 2019. https://www.britannica.com/technology/artificial-intelligence. Accessed 24 November 2019.

Bostrom, N. 2014. *Forms of superintelligence*. In Superintelligence: paths, dangers, strategies, Oxford: Oxford University Press.

Chassignol, M., A. Khoroshavin, A. Klimova, and A. Bilyatdinova. 2018. Artificial intelligence trends in education: a narrative overview. *Procedia Computer Science* 136: 16–24.

Chimuka, G. 2019. Impact of artificial intelligence on patent law. Towards a new analytical framework—[the Multi-Level Model]. *World Patent Information* 59: 101926.

Council of Europa (CE). 1997. Oviedo Convention: *Convention for the protection of human rights and dignity of the human being with regard to the application of biology and medicine: convention on human rights and biomedicine*. https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/164. Accessed 2 November 2019.

Dignum, V. 2018. Ethics in artificial intelligence introduction to the special issue. *Ethics and Information Technology* 20: 1–3.

Dorado-Diaz, P.I., J. Sampedro-Gomez, V. Vicente-Palacios, and P.L. Sanchez. 2019. Applications of artificial intelligence in cardiology; the future is already here. *Revista Espanola de Cardiologia*. https://doi.org/10.1016/j.rec.2019.05.014.

European Commission (EC). 2019. *EGTAI: the ethics guidelines for trustworthy artificial intelligence.*https://ec.europa.eu/futurium/en/ai-alliance-consultation. *Accessed 30 September 2019.*

European Commission (EC). 2018. High Level Expert Group on Artificial Intelligence. *A Definition of AI*. https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence. Accessed 25 November 2019.

Felzman, H., E.F. Villaronga, C. Lutz, and A. Tamo-Larrieux. 2019. Transparency you can trust: transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*. https://doi.org/10.1177/2053951719860542.

Future of Life*: Benefits & Risks of Artificial Intelligence.* 2016.https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/. *Accessed 24 November 2019.*

Future of Life. 2017a. *A Principled AI Discussion in Asilomar.* 2017a. https://futureoflife.org/2017/01/17/principled-ai-discussion-asilomar/. Accessed 25 November 2019.

Future of Life. 2017b. *AAIP: Asilomar AI Principles.* 2017b. https://futureoflife.org/ai-principles/. Accessed 30 September 2019.

Goering, S., and E. Yusta. 2016. On the necessity of ethical guidelines, for novel neurotechnologies. *Cell* 167: 882–885.

Guan, J. 2019. Artificial intelligence in healthcare and medicine: promises, ethical challenges and governance. *Chinese Medical Sciences Journal* 34: 76–83.

Haleem, A., D.M. Javaid, and I.H. Khan. 2019. Current status and applications of artificial intelligence (AI) in medical field: an overview. *Current Medicine Research and Practice*. https://doi.org/10.1016/j.cmrp.2019.11.005.

IEEE, Global Initiative on Ethics of Autonomous and Intelligent Systems. 2019. *Ethically aligned design.*https://ethicsinaction.ieee.org/#read. *Accessed 25 November 2019.*

Jiang, F., et al. 2017. Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology* 2: 230–243.

Jobin, A., M. Lenca, and E. Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1: 389–399.

Lee, Y., et al. 2019. Egoistic and altruistic motivation: how to induce users' willingness to help for imperfect AI. *Computers in Human Behavior* 101: 180–196.

Luxton, D.D., S.L. Anderson, and M. Anderson. 2016. Ethical issues and artificial intelligence technologies in behavioral and mental health care. *Artificial Intelligence in Behavioral and Mental Health Care*. https://doi.org/10.1016/B978-0-12-420248-1.00011-8.

Madannejada, R., et al. 2019. Toxicity of carbon-based nanomaterials: reviewing recent reports in medical and biological systems. *Chemico-Biological Interactions* 307: 206–222.

Marcus, G., and E. Davis. 2019. *Rebooting AI*. New York: Pantheon.

MDRDAI. 2018a. *Co-construction approach and methodology*. https://5dcfa4bd-f73a-4de5-94d8-c010ee777609.filesusr.com/ugd/ebc3a3_eb775ce43d3b46fe90c89da583e9744d.pdf Accessed 30 November 2019.

MDRDAI. 2018b. *Montréal declaration for a responsible development of artificial intelligence*. https://www.montrealdeclaration-responsibleai.com/ Accessed 30 September 2019.

Moor, J. 2006. The Dartmouth College artificial intelligence conference: the next fifty year. *AI Magazine* 27: 87–91.

Patle, B.K., et al. 2018. A review: on path planning strategies for navigation of mobile robot. *Defence Technology* 15: 582–606.

Say, C. 2018. *Yapay Zekanın Doğuşu. 50 Soruda Yapay Zeka*, p.83, Istanbul: 7 Renk Basın ve Yayın (in Turkish).

Simpson, B. 2016. Algorithms or advocacy: does the legal profession have a future in a digital world? *Information & Communication Technology Law* 25 (1): 50–61.

Tegmark, M. 2017. *Matter turns intelligent, In Life 3.0 Being human in the age of artificial intelligence*. New York: Alfred A. Knopf.

Turing, Alan. 1950. Computing Machinery and Intelligence. *Mind* 59: 433–460.

Wall, L.D. 2018. Some financial regulatory implications of artificial intelligence. *Journal of Economics and Business* 100: 55–63.

World Medical Association (WMA). 1964. Declaration of Helsinki—*Ethical Principles for Medical Research Involving Human Subjects*. https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/. Accessed 2 November 2019.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.