TECHNICAL AND ETHICAL RISKS

‘technical’ is an anagram of ‘nethical’

CONCRETE RISKS WITH FLEXIBLE SOLUTIONS

1. Avoiding **negative side effects**
2. Avoiding **reward hacking**
3. Ensuring **scalable oversight**
4. Ensuring **robustness to distributional shift**
5. Ensuring **safe exploration**

Amodei D, Olah C, Steinhardt J, *et al.* (2016) Concrete Problems in AI Safety. arXiv:1606.06565v2, pp 1–29. doi:1606.06565

1. AVOID NEGATIVE SIDE EFFECTS

1. Include an '**impact regularizer**' that penalizes change to the environment.
   - But how does the system represent change?
2. **Penalize influence**.
   - I.e., limit the amount/scope of resources available
   - But how does the system represent empowerment?
   - Do you penalize the AI if it *can* take an action, or if it *does*?

Amodei D, Olah C, Steinhardt J, *et al.* (2016) Concrete Problems in AI Safety. arXiv:1606.06565v2, pp 1–29. doi:1606.06565

2. AVOID REWARD HACKING

$$P(engagement|\text{📄}) = \lambda_1 P(Like|\text{📄}) + \lambda_2 P(comment|\text{📄}) + \cdots$$

1. **Abstract rewards**. Avoid the curse of dimensionality, especially with misbehaving numerical dimensions.
2. **Avoid Goodhart' Law** (*"when a metric is used as a target, it ceases to be a good metric"*).
   - E.g., avoid this logic: "if I give the patient lots of drugs, they'll stop coming into the office, ∴ **take all the drugs**!!!"

Amodei D, Olah C, Steinhardt J, *et al.* (2016) Concrete Problems in AI Safety. arXiv:1606.06565v2, pp 1–29. doi:1606.06565

3. SCALABLE OVERSIGHT & 4. DISTRIBUTIONAL SHIFT

3. A program trained on a few 100 examples might not generalize to an entire population.
   a) **Active learning** may help. Continuously rely on human consensus and input; validate 'difficult' data.
4. Can a system trained on *common* diseases capture *rare*, or *emergent* diseases?
   a) The system must **acknowledge** its own ignorance, and **resist** shifting its models too hastily.

Amodei D, Olah C, Steinhardt J, *et al.* (2016) Concrete Problems in AI Safety. arXiv:1606.06565v2, pp 1–29. doi:1606.06565

5. SAFE EXPLORATION

- Autonomous learning requires exploration, i.e., non-optimal actions which help the agent learn its environment.
  - **Bounded** or **simulated** exploration.
  - **Limit explorative influence** on distributions…

Amodei D, Olah C, Steinhardt J, *et al.* (2016) Concrete Problems in AI Safety. arXiv:1606.06565v2, pp 1–29. doi:1606.06565

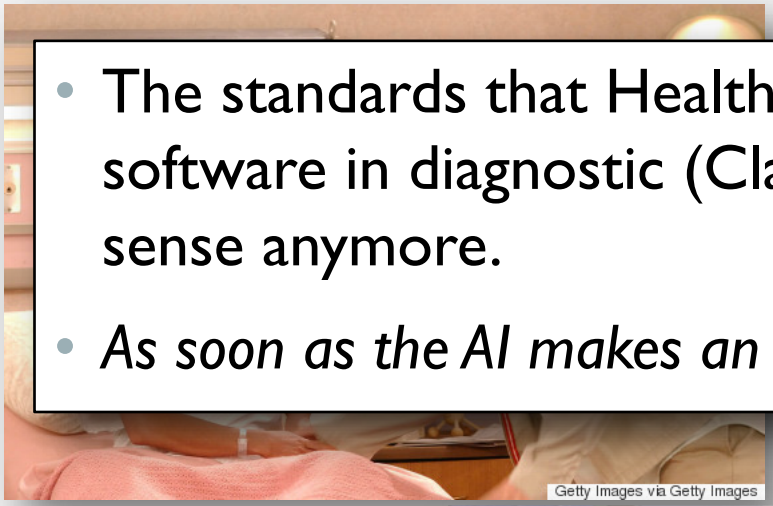
5. SAFE EXPLORATION & CLINICAL VS RESEARCH ETHICS

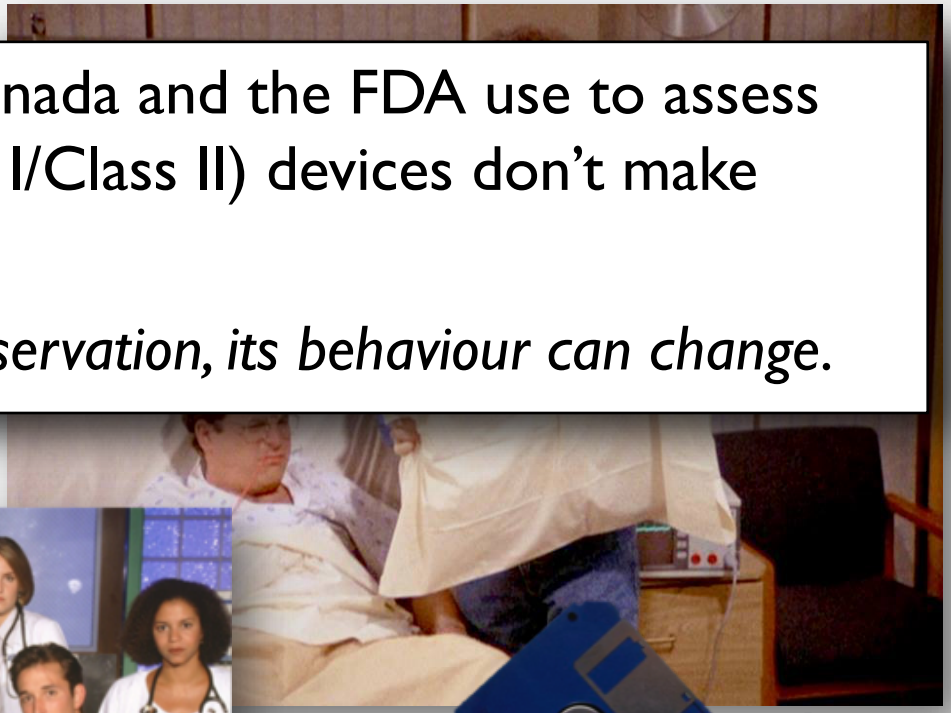
- A patient sees a doctor about a kind of depression that can be helped with anti-depressants, of which several are available.
  1. In the **Clinical** case, the doctor can prescribe drug X, after informing the patient of its benefits and side-effects.
  2. In the **Research** case, the doctor must obtain informed consent, and possibly prescribe drug Y, as an experiment, after explaining both drugs, the reason for their comparison, the randomness of prescription, and other obligatory details.

- If an 'AI doctor' is not only *prescribing* but also **directly** and **continuously learning** from outcomes, which set of ethics apply?

REGULATION OF MEDICAL DEVICES IS FROM THE 1990S

- The standards that HealthCanada and the FDA use to assess software in diagnostic (Class I/Class II) devices don't make sense anymore.
- *As soon as the AI makes an observation, its behaviour can change.*





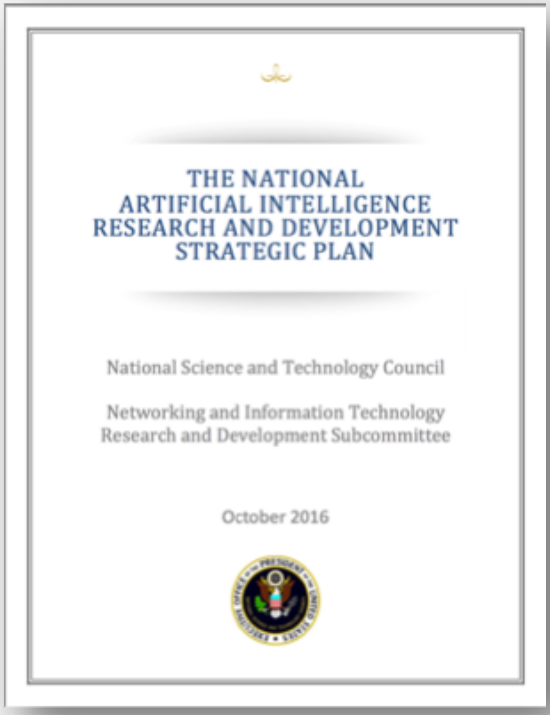
H.R.6 – 114TH CONGRESS 21ST CENTURY CURES ACT I

- The 21st Century Cures Act passed House of Representatives (344-77) 13 July 2015.
  - Received in the Senate, read twice, and referred to the Committee on Health, Education, Labor, and Pensions.
- **Guidance I**, "**general wellness products**": Include "audio recordings, video games, software programs and other products that are commonly … available from retail establishments."
  - The FDA will *not* regulate such products as medical devices, as long as they meet two factors, specifically they:
    i) are intended for only general wellness; and ii) present low risk to users.
  - These products' value derives from *information*, rather than doing something directly to the body.

John Graham, Artificial Intelligence, Machine Learning, And The FDA, 19 Aug 2016, Forbes,

STRATEGIES

- The Affordable Care Act shifted from a fee-for-service towards a pay-for-performance model[1]
  - Health IT is *rewarded*.
- Despite prohibitions in the Genetic Information Non-discrimination Act (2008), there is growing interest in using risk information for insurance stratification[2].
  - Differential pricing has become one of the standard practices for data analytics vendors, introducing new avenues to perpetuate inequality.
- **The (previous!) White House viewed AI as providing "increased medical efficacy, patient comfort, and less waste"[3].**



[1] David Blumenthal, Melinda Abrams, and Rachel Nuzum, "The Affordable Care Act at 5 Years," *New England Journal of Medicine* Vol. 372, Issue 25, (2015): 2453

[2] Yann Joly et al., "Life Insurance: Genomic Stratification and Risk Classification," European Journal of Human Genetics 22 No. 5 (May 2014): 575–79).

[3] Bryan Biegel, & Kurose, J. F. (2016). *The National Artificial Intelligence Research and Development Strategic Plan.*

THEY TOOK OUR JOBS!

THE GLOBE AND MAIL | Video

VIDEO HOME NEWS BUSINESS LIFE ARTS TECHNOLOGY DRIVE SPORTS

TECHNOLOGY Video: How artificial intelligence can help doctors diagnc

QUESTION

What is wrong with me ?

FACTORS

ars technica

MAIN MENU MY STORIES: 25 FORUMS SUBSCRIBE JOBS

SCIENTIFIC METHOD / SCIENCE & EXPLORATION

Robots: Destroying jobs, our economy, and possibly the world

FORTUNE

SUBSCRIBE

Technology will replace % of what doctors do

hosla DECEMBER 4, 2012, 2:26 PM EST

WIRED

Artificial Intelligence Is Now Telling Doctors How to Treat You

BUSINESS CULTURE DESIGN GEAR SCIENCE

DANIELA HERNANDEZ, KAISER HEALTH NEWS BUSINESS 06.02.14 6:30 AM

SHARE

SHARE 2693

TWEET

PIN 159

COMMENT 22

EMAIL

ARTIFICIAL INTELLIGENCE IS NOW TELLING DOCTORS HOW TO TREAT YOU

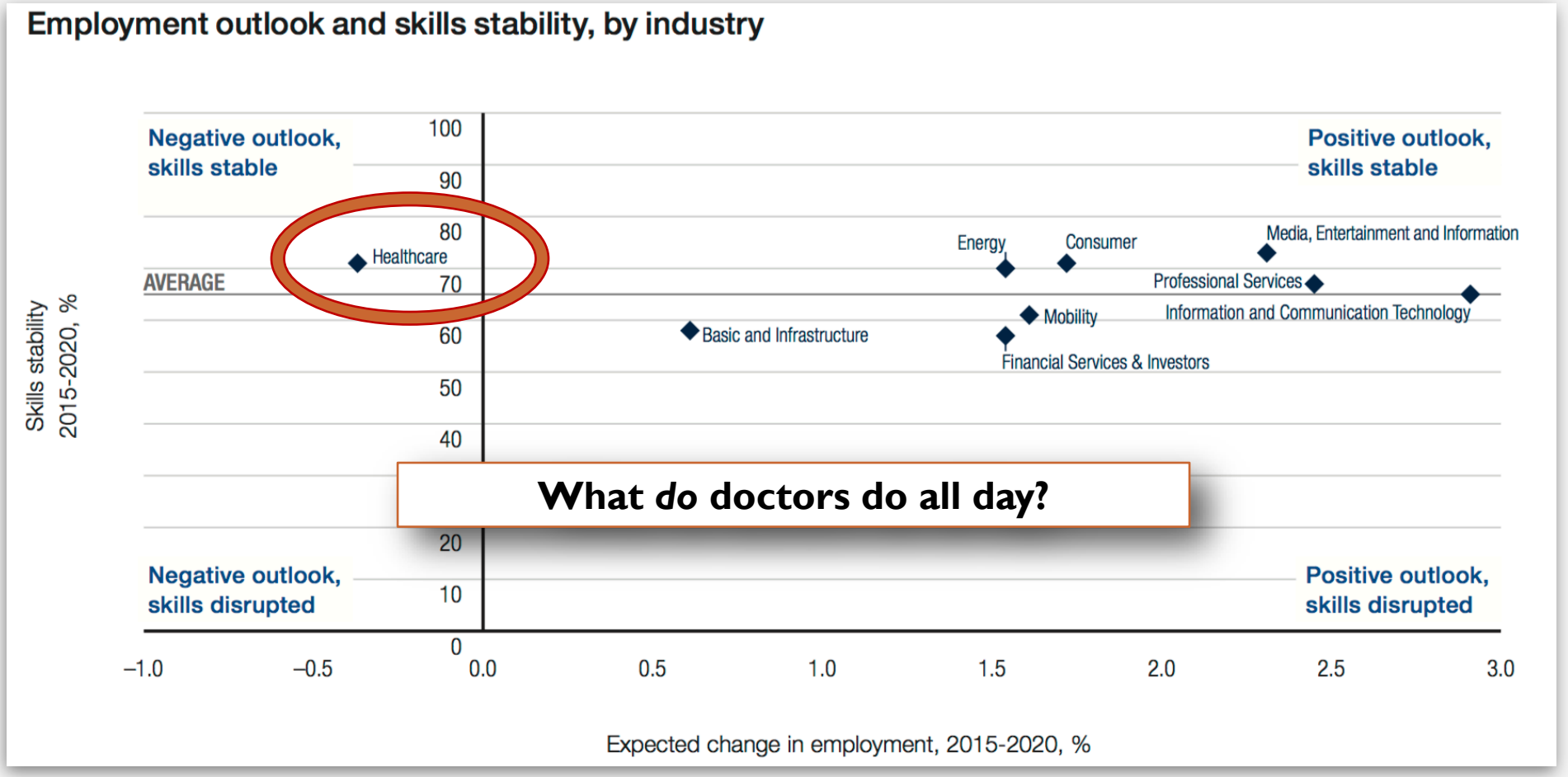
subscribe search dating more International

theguardian

sport football opinion culture business lifestyle fashion environment tech travel all sections

Would you bet against sex robots? AI 'could leave half of world unemployed'

Scientist Moshe Vardi tells colleagues that change could come within 30 years, with few professions immune to effect of advanced artificial intelligence



"Technology will replace 80% of what doctors do"
– Vinod Khosla

The World Economic Forum (2016) "The Future of Jobs Employment. Skills and Workforce Strategy for the 4th Industrial Revolution",

HUMANS MAKE MISTAKES

- **Humans** are notoriously **bad** with **information**.
  - Patients **misread** or **miscommunicate** their own symptoms.
    - Nearly **half** of American adults have difficulty understanding and acting upon health information (IOM, 2004).
  - Faulty memory; skill obsolescence; cognitive biases; cognitive/time limitations; **recency biases**; other human biases.
    - *Diagnoses* **correlate** with *advertising* and *media* exposure.
- Winters *et al.* (2012) showed that ~40,500 patients die in ICU, in the USA, each year due to misdiagnosis.

http://www.nap.edu/openbook.php?record_id=10883&page=1

Winters *et al.* (2012) Diagnostic errors in the intensive care unit: a systematic review of autopsy studies. *BMJ Qual Saf* 2012;**21**:894-902

HUMANS MAKE MISTAKES

- Graber *et al.* (2005) studied one hundred cases of **diagnostic error** involving internists …
  - **Cognitive factors** contributed to 74% of cases.
    - Most common cause: 'premature closure'.
- Eddy (1990) showed top surgeons descriptions of surgical problems and asked: *Should the patient have surgery?*
  - 50% said **Yes**, 50% said **No**.
  - 40% gave conflicting answers upon retesting.

Graber et al. (2005) Diagnostic Error in Internal Medicine. *Arch Intern Med.*, **165**(13):1493-1499
Eddy (1990) The Challenge. *JAMA*, **263**(2):287-290. http://jama.jamanetwork.com/article.aspx?articleid=380215

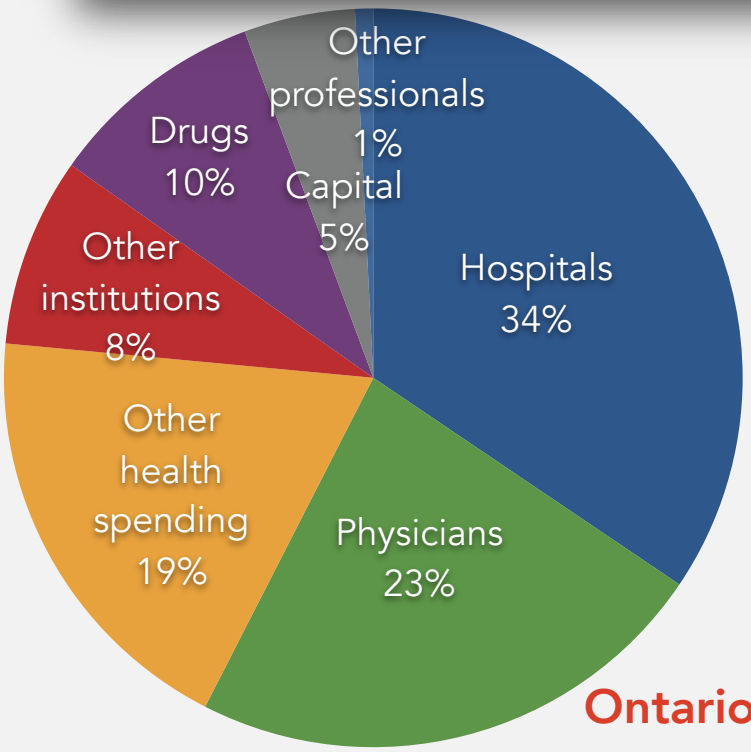
AI TO THE RESCUE?

- Bennett and Hauser (2013) compared patient outcomes between doctors and *sequential decision-making algorithms* using 500 randomly selected patients.
  - Estimated AI cost: $189; Human cost: $497.
  - Outcomes up to 50% better using AI.

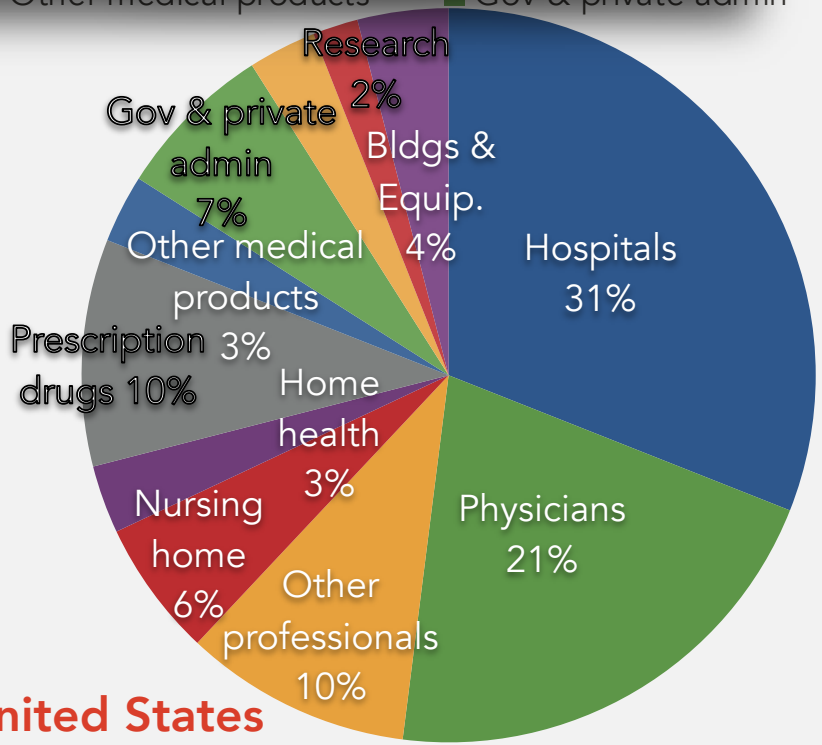
Bennett and Hauser (2013) Artificial intelligence framework for simulating clinical decision-making: a Markov decision process approach. Artif Intell Med. 57(1):9-19

WHERE WILL CHANGE HAPPEN?

"from a solely hospital-centred system [towards] a community [primary care] system"
Premier Kathleen Wynne, 2017



Ontario



United States

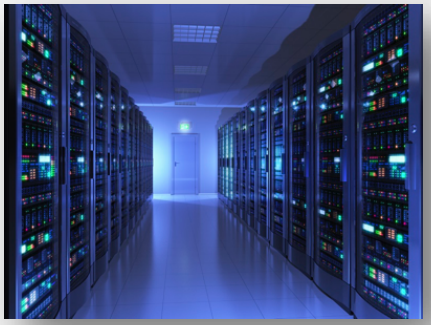
National Health Expenditure Trends, Canadian Institute for Health Information, 2010

Department of Health and Human Services, 2011

FINAL THOUGHTS

THE QUANTIFIED SELF VS THE MEDICAL RECORD

- Many apps serve to **shift** the **responsibility** for care and monitoring from healthcare professionals to patients themselves.
  - This may disadvantage patients who do not have the time, resources, or access to technology.
  - **What kinds of patients are favored in this new dynamic**, and might patients not well-equipped to manage and maintain their own data receive substandard care?
  - What new roles and responsibilities do the *developers* of such apps take on, and how do the ethical responsibilities of medical professionals get integrated into these differing contexts?.
- How to combine *models* in different AIs? There's no EDI in HIPAA for *models*.



VS


Crawford, K., Whittaker, M., Elish, M. C., Barocas, S., Plasek, A., & Ferryman, K. (2016). *The AI Now Report.*

CHANGING OUR OBJECTIVE FUNCTIONS

- Regulatory changes need to continue to **respect** individual rights.
- But what if the spectre of **surveillance capitalism** can actually *help* the individual?
  - What good is an AI agent that can only learn on your few dozen EMR records, vs one that can learn from *millions*.
  - If the potential of Big Data is not met, patients will not benefit.

CHANGING OUR OBJECTIVE FUNCTIONS

- There is movement – but are we ready?
  - Corporate EMR
  - Nascent partnership between ICES and Vector
  - Quantified self and **AI startups**
- How to promote economic growth through innovative health sector spending in a public system?
- How to balance **population** medicine with **patient-centred** care, in an AI sense?

TRENDING NOW (REDUX)

1. Deep neural networks (of course)
2. Big Data (with cells interlinked within cells interlinked)
3. Recurrent neural networks for temporal, dynamic data
4. Reinforcement learning
5. Active learning
6. Telehealth and remote monitoring
7. Causal, explainable models

Who **accesses** the data? How **accurate** must these systems be?
How are **costs** weighed against **outcomes**? Who is **liable**?

The word "diagnosis," … comes from the Greek for "**knowing apart**." Machine-learning algorithms will only become better at such knowing apart—at partitioning, at distinguishing moles from melanomas. But knowing, in all its dimensions, transcends those task-focussed algorithms. In the realm of medicine, perhaps the ultimate rewards come from **knowing together**.

Siddartha Mukherjee (2017),
AI Versus MD, *The New Yorker,* 3 April