Elsevier Editorial System(tm) for The Lancet
Manuscript Draft

Manuscript Number: THELANCET-D-19-05594R1

Title: The Cognitive Health System

Article Type: Viewpoint

Corresponding Author: Professor Enrico Coiera,

Corresponding Author's Institution:

First Author: Enrico Coiera

Order of Authors: Enrico Coiera

The Cognitive Health System

Enrico Coiera MB BS PhD
Australian Institute of Health Innovation, Macquarie University, level 6 75 Talavera Rd, NSW 2109, Australia.
enrico.coiera@mq.edu.au
Tele: +61 437044180

Word count: 1850 excluding table and box
References: 22

“What if the health system could think?” Turing once famously asked a similar question about machines[1]. Now artificial intelligence (AI) challenges us to reimagine medicine in a machine-assisted world [2]. But what will that world be like? It is easy to focus just on machines and what they can do, but healthcare is not a software problem to be solved by algorithms and machine learning alone. Humans populate our sociotechnical system, and we both create technology, and in turn are shaped by it - technical systems have social consequences, and social systems have technical consequences [3].

As our healthcare services are digitized, interconnected and automated, the nature of the health system necessarily will transform. *Distributed cognition* researchers have long held that human reasoning is situated, meaning it is a product both of the reasoner and the wider world they inhabit [4-6]. Humans decisions are often opportunistic, influenced by the information right in front of us, and by the people or tools (such as guidelines or computers) that can help with our reasoning. In any organisation, decisions emerge not from isolated internal processes within the brains of humans or the code of machines, but from the many distributed interactions that occur between these.

To this distributed cognitive system that we call healthcare, we now add AI tools that directly influence human decisions and will increasingly have autonomy to make decisions. Such a distributed network of humans and AI is called a *cybersocial* system, with actions and outcomes emerging from the interacting decisions of both human and machine[7-9]. While a cybersocial health system might not exactly “think” on its own, it will exhibit new adaptive and emergent behaviors, giving rise to what we might call the cognitive health system.

The nature of cybersocial systems

Today, because such ideas may seem the stuff of science fiction, we remain singularly unprepared to either exploit the potential of such a world, or capable of managing its unintended consequences. It is one thing to imagine a learning health system which exploits past lessons to improve future care [10], it is another to inhabit a system which behaves unpredictably or non-transparently.

Yet cybersocial systems are already commonplace. Today’s social media platforms create worlds where it can be hard to distinguish human from ‘bot’, and where information is algorithmically filtered and prioritized. It is the nature of a system like Facebook that global behaviors are adaptive and emergent, the joint product of human and machine. The system can be nudged and cajoled by those with an agenda, but even its creators do not have the ability to fully control or predict how it will behave.

The presence of AIs in a distributed cognitive system brings new challenges to humans who would seek to 'control' it (Box 1). Even simple AI enabled software agents are able to be adaptive, and can carry out transactions such as sending messages with a velocity and volume that can far exceed the capacity of humans to engage. A coalition of humans and machine could thus use machine learning to first identify the most effective strategy to meet a goal, and then trigger a surge of transactions that are effectively unmanageable by humans. We have already seen this behavior in surges of "fake" political messaging on social media in the immediate run up to elections that cannot be verified in the time available. In such a case, verification delayed becomes verification denied. The degree to which humans can in any sense 'control' a cybersocial system thus depends on the nature of these human-machine coalitions and specifically on the capability of humans to manage the volume and type of machine transactions that are generated.

Box 1: Attributes of a cyber-social system

Adaptation: Machine learning capabilities allow software agents to identify those actions that maximise the chance of achieving their goals. Trialling different strategies on human or other machine agents provides data on optimal strategy.

Volume and Velocity: Software agents can generate super-human numbers of transactions, at super-human rates, such as sending messages or searching though data.

Cloakabilty: Software agents may be able to mask their identity by assuming the outward attributes of other human or software agents.

Clonabilty: A given software agent can rapidly duplicate itself, creating a large number of identical agents; agent homogeneity amplifies the effects of the agent type within a network.

Coalitions: A cybersocial system will be composed of a number of potentially overlapping human and machine coalitions. Coalitions may act internally with common purpose while competing with other coalitions. The ratio of human to machine in a coalition will impact on possible behaviours and susceptibilities.

Surges: Software enabled coalitions can create transaction surges, such as creating a high volume of messages over a short time period, that exceed the capability of humans to handle them.

Asymmetric verifiability: Transaction surges become impossible to verify by humans, potentially leading to outcomes not otherwise be possible with a more balanced human capability to engage.

Challenges and opportunities and for healthcare

Public and population health

Cybersocial systems already impact population health, for example by amplifying anti-vaccination messages, and sharing them in algorithmically targeted ways across friendship networks. Machine learned algorithms encourage friendship network building by prioritizing message sharing over message safety. The resulting online contagion of false beliefs leads to real world contagion of infectious disease, as individuals alter behaviors and avoid vaccination[11]. More worryingly, social media has been weaponized by malicious actors in the political sphere, including the targeting of anti-vaccination messages [12]. Public health strategies now have to recognize the risks of online manipulation of human behaviour. One can even imagine how a maliciously created surge of online messages about a "fake" disease outbreak could trigger widespread panic, swamped health services, and hoarding behaviour that leads to shortages of water, food and medicines.

To counter such risks, real-time access to cybersocial data can lead to rapid and precise public health responses to disease outbreaks, real or "fake". Community acquired data such as a person's location, movement and online engagements, ethically and consensually obtained, can help both in tracking disease, monitoring emerging population behavior and beliefs, and shape health services responses. Distributed AI can assist by delivering public health messages that are customized at the community or individual level, with a speed and reach not possible in a less adaptive system.

Primary care

New patient facing technologies can also have unintended social consequences. AI driven symptom checkers for example provide patients with advice on whether or not to seek medical care, or visit a hospital emergency room[13]. If these technologies become widely used as part of a 'digital first' primary care service, then there is the potential for large-scale behaviour change that significantly alters service demand. Positive changes might include reducing health service loads by keeping patients with seasonal viral illnesses at home and off antibiotics. Accidental or intentional flaws in this algorithmic advice may have negative impacts, for example unnecessarily encouraging patients to seek help from already stretched health services.

A cybersocial learning health system

A learning health system is one that can quickly identify new evidence from the experiences of patients, as captured in routine clinical data, and translate that evidence into systemic changes in practice[14]. From a technical viewpoint, such a system will be built from massive clinical datasets, machine learning, and computable models of clinical knowledge. Yet a learning health system is also obviously cybersocial because evidence creation, dissemination and adoption are in part socially shaped[9]. Research is contested, conflicts of interest and biases shape evidence production and synthesis, and clinical recommendations are never uniformly adopted.

The great potential of a learning health system is the mechanism it provides to accelerate the discovery of new treatments, workflows, and personalization of care. Connecting health services with the broader community could also contribute to system adaptation beyond healthcare. Long term planning of urban geography, transport and energy, and the volume and type of health services in a community can all be informed dynamically by the thousands of digital behavioural experiments occurring across the population. Ensuring such machine adaptations do not reinforce social disparities by modelling pre-existing biases and the digital divide will be an ongoing challenge.

Increasing the speed of evidence production and adoption is a tantalising goal, but it may have unexpected cybersocial consequences. Managing change in the healthcare system has long been a challenge of juggling limited resources, sunk costs and vested interests. Faster shifts to new best practices, enabled by machine learning, also means faster obsolescence of past practices, challenging our ability to plan for and manage change. Stockpiles of medicines could sit unwanted, resulting not just in waste, but economic cost for manufacturers. Shortages would arise if there is no capacity to ramp up production to meet unexpected demand. Professionals might find the demand for their skills dissolve or quickly become greater than they can manage. Those with most to lose are likely to resist change, as we see with some industry responses to climate change. Accelerated adoption of new practices may also see traditional safety checks side-lined in the pursuit of earlier adoption or profit.

Harnessing the cognitive health system

Today we are well equipped to respond to many local health system challenges, but struggle with whole of system responses to problems like health system resilience, sustainability and safety. Creating a cognitive health system might just do that for us. If well crafted, this new system will exhibit responses that do not always require top-down directives. Rather, they can emerge from the collective interaction of AI and human.

At this moment however, we are largely unequipped to either exploit or manage the cybersocial nature of our emerging smart and networked world. Our ambitions for the construction of a learning health system are still embryonic. Our understanding of the way blended AI-human collectives function is just as unevolved and requires interdisciplinary study[15]. There is a rich but not well integrated literature on the emergence of behaviour in social systems spanning sociology, economics, cognitive science, complex systems, network theory and AI [16]. The challenge for researchers is to harness this multidisciplinarity and to learn from the many natural cybersocial experiments already underway. For example, by modelling the effects of automated bots and opinion leaders in social networks, we should be in a position to explain emergent system behaviours such as vaccine refusal or the polarisation of political views [17,18]. Seeing cybersocial systems as “living” organisms reveals high level dynamic patterns of system growth, transformation, and even death [19].

When cognition is distributed, so is agency and accountability. We must work to develop new strategies that can manage and exploit the unique characteristics of a cybersocial world. As we see with platforms such as Facebook, governing either individual or system behaviour is a challenging and moving target. Interventions that worked once can quickly become obsolete as new behaviours emerge, either from intentional workarounds, or as a natural evolution.

Strategies to create a system that behaves in the way we wish could take a number of approaches (Table 1). The ‘surprise’ of emergent system behaviour is often related to the existence of unexpected connections and the way they allow new interactions and outcomes. Surprises may take the form of novel behaviour, or interestingly, in the form of inertia to planned change[20]. The more hidden or ‘dark’ complexity there is in a system, the less likely we are to be able to predict its behaviour. Whenever new local health system interventions are attempted, it will be critical to not just assess their proximate impact, but also to model the connections of the local intervention to the broader system and identify unexpected system consequences.

The research agenda to understand and exploit cybersocial systems presents a formidable challenge to healthcare as we do not routinely study our system from the perspective of complex networks, we do not think about the behavior of human-machine coalitions, and we are far more comfortable with targeted health service interventions. If we get this right, we will witness the birth of new working partnerships between humans and AI, new social dynamics, and new pathways to system change. The opportunity for healthcare is to conceive of a vastly different way to share and shape knowledge at scale, and to craft new, more connected ways of delivering care and living our lives. While our new and distributed cognitive health system might not think in the way we traditionally understand it, it will surely have a life of its own[21].

Table 1: Cybersocial challenges and potential approaches to their governance in healthcare.

Characteristic	**Implication**	**Management strategy**
Dark Complexity: Much system complexity, originating out of interactions between humans and machines is unobserved. Whether because it is unmonitored, unmapped, or unmeasurable, hidden complexity does not reveal itself until specific circumstances arise.	Responses to system interventions are inherently unpredictable. Definitive mapping of system components and all their interactions is not possible.	Localized mapping of safety critical system elements is essential, and testing of interventions in a wide variety of contexts and scenarios should minimize the risk of surprise. When new system elements are constructed, controlling the points at which they interact within the system will minimize unexpected interactions.
Redundant information flows: Complex interconnection makes it is hard to control information flows. Disintermediation (the removal of "middle men") is always possible where alternate information flows can be constructed around those who 'control' information flow.	Trying to control system behaviours by controlling information flow is easily outmanoeuvred. There are always opportunities for informational attacks such as message distortion or information overload.	Redundancy in information strategies decreases risks of being side-lined or attacked. "Innoculation" strategies may improve the resilience of decision makers to informational attacks. Distributed decision making by consensus may minimise the impact of localised attacks. Systems can be designed to help cope with unwanted or malicious information surges.
Rapid adaptation: Embedding distributed AI and human decision making into a complex adaptive system has the potential to make system adaptation unfold over a short timescale.	Early responses to system level shocks may minimize their impact and improve system resilience. Rapid change in best practice recommendations may improve the quality of clinical outcomes. Changes in best practice may lead to rapid and expensive obsolescence of services, technologies and therapies. Unwarranted adaptation may lead to waste and cause harm.	Close coupling of health service adaptation to real world outcomes, similar to post-market surveillance of medications, should help to minimise unwarranted adaptation or unexpected problems.

Emergent decision making: The actions of individual human and machine agents to satisfy local goals may collectively have unexpected impacts globally e.g. stock market crashes from high speed micro-trading.	Local behaviours can have global implications. Changes in the behavior of an AI external to but connected to your local system may result in changes to local decision-making.	Local actors need to be aware of potential global effects and develop strategies that manage these. This includes monitoring external dependencies and vulnerabilities, and unexpected changes in local decisions. Cognitive engineering, the shaping of network dependencies for decision-making may increase transparency and resilience.
Uncontrolled increase in complexity: Without system-wide governance, there is no inherent mechanism to reduce complexity.	System characteristics such as inertia to change become more pronounced with time.	System 'apoptosis' mechanisms relying on AI can be envisaged from local (delete a guideline that is no longer accurate) to the global (decommission infrastructure or services that are not state of art) to create a 'forgetting health system'[22].
A society of systems: The health system sits alongside and interacts with many other systems including transport, education, finance etc.	Goals of different systems may conflict, neutralizing intended outcomes in one system, and triggering unwanted outcomes in others.	Explicit engineering of interfaces between critical systems, and explicit planning for managing goal conflicts between systems.

References

1. Turing AM. Computing machinery and intelligence. Mind 1950:433-60.
2. Coiera E. The fate of medicine in the time of AI. The Lancet 2018;392:2331-2.
3. Coiera E. Four rules for the reinvention of health care. British Medical Journal 2004;328:1197-9.
4. Hollan J, Hutchins E, Kirsh D. Distributed cognition: toward a new foundation for human-computer interaction research. ACM Transactions on Computer-Human Interaction (TOCHI) 2000;7:174-96.
5. Furniss D, Garfield S, Husson F, Blandford A, Franklin B. Distributed Cognition: Understanding Complex Sociotechnical Informatics. Studies in health technology and informatics 2019;263:75-86.
6. Suchman L. Human-Machine Reconfigurations: Plans and Situated Actions Boston MA: Cambridge University Press; 2006.
7. Stefanelli M. The socio-organizational age of artificial intelligence in medicine. Artificial intelligence in medicine 2001;23:25-47.
8. Wang F-Y. The emergence of intelligent enterprises: From CPS to CPSS. IEEE Intelligent Systems 2010;25:85-8.
9. Friedman C, Rubin J, Brown J, et al. Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. Journal of the American Medical Informatics Association 2014;22:43-50.
10. Friedman CP, Wong AK, Blumenthal D. Achieving a Nationwide Learning Health System. Science Translational Medicine 2010;2:57cm29-57cm29.
11. Dunn AG, Leask J, Zhou X, Mandl KD, Coiera E. Associations between exposure to and expression of negative opinions about human papillomavirus vaccines on social media: an observational study. Journal of medical Internet research 2015;17:e144.
12. Broniatowski DA, Jamison AM, Qi S, et al. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. American journal of public health 2018;108:1378-84.
13. Fraser H, Coiera E, Wong D. Safety of patient-facing digital symptom checkers. The Lancet 2018;392:2263-4.
14. Friedman C, Rubin J, Brown J, et al. Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. Journal of the American Medical Informatics Association : JAMIA 2015;22:43-50.
15. Rahwan I, Cebrian M, Obradovich N, et al. Machine behaviour. Nature 2019;568:477.
16. Gilbert N. Emergence in social simulation. Artificial societies: Routledge; 2006:134-43.
17. Bond RM, Fariss CJ, Jones JJ, et al. A 61-million-person experiment in social influence and political mobilization. Nature 2012;489:295-8.
18. Dunn AG, Surian D, Leask J, Dey A, Mandl KD, Coiera E. Mapping information exposure on social media to explain differences in HPV vaccine coverage in the United States. Vaccine 2017;35:3033-40.
19. Zheng M, Cao Z, Vorobyeva Y, Manrique P, Song C, Johnson NF. Multiscale dynamical network mechanisms underlying aging of an online organism from birth to death. Scientific Reports 2018;8:3552.
20. Coiera E. Why system inertia makes health reform so difficult. British Medical Journal 2011;342.

21. Barabási A-L. Linked: The new science of networks. AAPT; 2003.
22. Coiera E. The forgetting health system. Learning Health Systems 2017;1:e10023.

*Manuscript with revisions highlighted

The Cognitive Health System

Enrico Coiera MB BS PhD
Australian Institute of Health Innovation, Macquarie University, level 6 75 Talavera Rd, NSW 2109, Australia.
enrico.coiera@mq.edu.au
Tele: +61 437044180

Word count: ~~1603~~1850 excluding table and box
References: ~~20~~22

"What if the health system could think?" Turing once famously asked a similar question about machines[1]. Now artificial intelligence (AI) challenges us to reimagine medicine in a machine-assisted world [2]. But what will that world be like? It is easy to focus just on machines and what they can do, but healthcare is not a software problem to be solved ~~just~~ by algorithms and machine learning alone. Humans populate our sociotechnical system, and we both create technology, and in turn are shaped by it - technical systems have social consequences, and social systems have technical consequences [3].

As our healthcare services are digitized, interconnected and automated, the nature of the health system necessarily will transform. ~~*Instead of cognition being an internal process residing within the brains of humans or the code of machines, cognition will increasingly happen at a system-level.*~~ *Distributed cognition* researchers have long held that human ~~cognitive tasks are~~reasoning is situated, meaning it is a product both of the reasoner and the wider world they inhabit ~~so tightly coupled to the environment that cognition necessarily extends into the world~~ [4-6]. ~~In complex environments, h~~Humans decisions are often opportunistic, influenced by the information right in front of us, and by the people or tools~~necessarily delegate elements of reasoning to tools~~ (such as ~~clinical~~ guidelines or computers) that can help with our reasoning ~~or to other humans~~. In any organisation, decisions emerge not from isolated internal processes within the brains of humans or the code of machines, but from the many distributed interactions that occur between these.
~~Instead of cognition being an internal process residing within the brains of humans or the code of machines, cognition will increasingly happen at a system-level.~~

To this distributed cognitive system that we call healthcare, we now add AI ~~systems are~~ tools ~~with a difference in~~ that directly influence human decisions ~~they~~ and will increasingly ~~will~~ have ~~agency~~ autonomy to make decisions~~, or directly influence human decisions~~. Such ~~Our adoption of AI is thus creating a new kind of distributed cognitive system, one that is characterized by a~~ a ~~collaborative~~ distributed network of humans and AI~~[6,7]~~ is called a~~.~~ ~~In this~~ *cybersocial* system, with actions and outcomes emerg~~ing~~e ~~out of~~from the ~~collective~~ interacting decisions of both human and machine[7-9]. While ~~this~~ a cybersocial health system might not exactly ~~actually~~ "think" on its own, it will exhibit new adaptive and emergent behaviors ~~that are emergent, adaptive and for practical purposes a form of reasoning~~, giving rise to what we might call the~~a~~ cognitive health system.

Formatted: Font: Italic

Formatted: Font: Font color: Auto, English (Australia)

The nature of c~~C~~ybersocial systems ~~bring new opportunities and risks for healthcare~~

Today, because such ideas may seem the stuff of science fiction, we remain singularly unprepared to either exploit the potential of such a world, or capable of managing its unintended consequences. It is one thing to imagine a learning health system which exploits past lessons to improve future care [10], it is another to inhabit a system which behaves unpredictably or non-transparently.

Yet cybersocial systems are ~~not creatures of fantasy – indeed they are~~already commonplace. ~~Wikipedia, an online collective authoring system, is a prime example of a distributed cognitive system, facilitated by collaborative computational tools [10,11].~~ Today's social media platforms ~~go much further,~~ create~~ing~~ worlds where it can be hard to distinguish human from 'bot', and where information is algorithmically filtered and prioritized. It is the nature of a system like Facebook that global behaviors are adaptive and emergent, the joint product of human and machine. ~~It~~ The system can be nudged and cajoled by those with an agenda, but even its creators do not have the ability to fully control or predict how it will behave.

The presence of AIs in a distributed cognitive system brings new challenges to humans who would seek to 'control' it (Box 1). Even simple AI enabled software agents are able to be adaptive, and can carry out transactions such as sending messages with a velocity and volume that can far exceed the capacity of humans to engage. A coalition of humans and machine could thus use machine learning to first identify the most effective strategy to meet a goal, and then trigger a surge of transactions that are effectively unmanageable by humans. We have already seen this behavior in surges of "fake" political messaging on social media in the immediate run up to elections that cannot be verified in the time available. In such a case, verification delayed becomes verification denied. The degree to which humans can in any sense 'control' a cybersocial system thus depends on the nature of these human-machine coalitions and specifically on the capability of humans to manage the volume and type of machine transactions that are generated.

Box 1: Attributes of a cyber-social system

Adaptation: Machine learning capabilities allow software agents to identify those actions that maximise the chance of achieving their goals. Trialling different strategies on human or other machine agents provides data on optimal strategy.

Formatted: Justified

Formatted: Font: 14 pt

Formatted: Font: 14 pt

Formatted: Font: 14 pt

Formatted: Font: 14 pt

Formatted: Font: 14 pt

Formatted: Font: 14 pt

Formatted: Font: 14 pt

Formatted: Font: 14 pt

Formatted: Font: 14 pt

Formatted: Font: 14 pt

Formatted: Font: 14 pt

Formatted: Centered, Border: Left: (Single solid line, Auto, 0.5 pt Line width, From text: 1 pt Border spacing:), Right: (Single solid line, Auto, 0.5 pt Line width, From text: 1 pt Border spacing:)

Volume and Velocity: Software agents can generate super-human numbers of transactions, at super-human rates, such as sending messages or searching though data.

Cloakabilty: Software agents may be able to mask their identity by assuming the outward attributes of other human or software agents.

Clonabilty: A given software agent can rapidly duplicate itself, creating a large number of identical agents; agent homogeneity amplifies the effects of the agent type within a network.

Coalitions: A cybersocial system will be composed of a number of potentially overlapping human and machine coalitions. Coalitions may act internally with common purpose while competing with other coalitions. The ratio of human to machine in a coalition will impact on possible behaviours and susceptibilities.

Surges: Software enabled coalitions can create transaction surges, such as creating a high volume of messages over a short time period, that exceed the capability of humans to handle them.

Asymmetric verifiability: Transaction surges become impossible to verify by humans, potentially leading to outcomes not otherwise be possible with a more balanced human capability to engage.

Challenges and opportunities and for healthcare

Public and population health

Cybersocial systems already impact population health, for example by amplifying anti-vaccination messages, and sharing them in algorithmically targeted ways across friendship networks. Machine learned algorithms encourage friendship network building by prioritizing message sharing over message safety. The resulting online contagion of false beliefs leads to real world contagion of infectious disease, as individuals alter behaviors and avoid vaccination[11]. More worryingly, social media has been weaponized by malicious actors in the political sphere, including the targeting of anti-vaccination messages [12]. Public health strategies now have to recognize the risks of online manipulation of human behaviour. One can even imagine how a maliciously created surge of online messages about a "fake" disease outbreak~~s~~ could trigger widespread ~~real world~~ panic, swamped health services, and hoarding behaviour that ~~leading~~ leads to shortages of water, food and medicines~~, and swamped health services~~.

To counter such risks, real-time access to cybersocial data can lead to rapid and precise public health responses to disease outbreaks, real or

Formatted: Font: Bold

Formatted: Font: Bold

Formatted: Font: Bold

Formatted: Font: Italic

"fake". Community acquired data such as a person's location, movement and online engagements, ethically and consensually obtained, can help both in tracking disease, monitoring emerging population behavior and beliefs, and shape health services responses. Distributed AI can assist by delivering public health messages that are customized at the community or individual level, with a speed and reach not possible in a less adaptive system.

Primary care

Formatted: Font: Italic

New patient facing technologies can also have unintended social consequences. AI driven symptom checkers for example provide patients with advice on whether or not to seek medical care~~-~~, or visit a hospital emergency room[13]. ~~Accidental or intentional flaws in this algorithmic advice will affect whether or not patients seek help.~~ ~~In some circumstances~~If these technologies become widely used as part of a 'digital first' primary care service, -then ~~we may see~~there is the potential for large-scale behaviour change that significantly ~~distorts~~ alters service demand. Positive changes might include reducing health service loads by keeping patients with seasonal viral illnesses at home and off antibiotics. Accidental or intentional flaws in this algorithmic advice ~~will affect whether or not patients seek help.~~may have negative impacts, for example unnecessarily encouraging patients to seek help from ~~and overloads~~ already stretched health services.

~~To counter such risks, real-time access to cybersocial data can lead to rapid and precise public health responses to disease outbreaks, real or "fake". Community acquired data such as a person's location, movement and online engagements can help both in tracking disease, monitoring emerging population behavior and beliefs, and shape health services responses. Distributed AI can assist by delivering messages that are customized at the community or individual level, with a speed and reach not possible in a less adaptive system.~~

Formatted: English (United Kingdom)

A cybersocial learning health system

Formatted: Font: Not Bold, Italic

Formatted: Font: Italic

A learning health system is one that can quickly identify new evidence from the experiences of patients, as captured in routine clinical data, and translate that evidence into systemic changes in practice[14]. From a technical viewpoint, ~~the learning health~~such a system will be built from massive clinical datasets, machine learning, and computable models of clinical knowledge. Yet a learning health system is also obviously cybersocial because evidence creation, dissemination and adoption are in part socially shaped[9]. Research is contested, conflicts of interest and biases shape evidence production and synthesis, and clinical recommendations are never uniformly adopted.

The great potential of a learning health system is the mechanism it provides to accelerate the discovery of new treatments, workflows, and personalization of care. Connecting health services with the broader community ~~can~~ could also contribute to ~~broader~~ system adaptation beyond healthcare. Long term planning of urban geography, transport and energy, and the volume and type of health services in a community can all be informed dynamically by the thousands of digital behavioural experiments occurring across the population. Ensuring such machine adaptations do not reinforce social disparities by modelling pre-existing biases and the digital divide will be an ongoing challenge.

~~Yet the consequences of i~~Increasing the speed of evidence production and adoption is a tantalising goal, but it ~~will also~~may have unexpected cybersocial consequences. ~~For example, m~~Managing change in the healthcare system has long been a challenge of juggling limited resources, sunk costs and vested interests. Faster shifts to new best practices, enabled by machine learning~~ and computational discovery systems~~, ~~will~~ also means faster obsolescence of past practices~~. Yet a rapid shift in the demand for particular health services may~~, challenging~~e~~ our ability to plan for and manage change. Stockpiles of medicines could sit unwanted, resulting not just in waste, but economic cost for manufacturers. Shortages would arise if there is no capacity to ramp up production to meet unexpected demand. Professionals might find the demand for their skills dissolve or quickly become greater than they can manage. Those with most to lose are likely to resist change, as we ~~are~~ see~~ing~~ with some industry responses to climate change. Accelerated adoption of new practices may also see traditional safety checks side-lined in the pursuit of earlier adoption or profit.

Harnessing the cognitive health system

Today we are well equipped to respond to many local health system challenges, but struggle with whole of system responses to problems like health system resilience, sustainability and safety. Creating a cognitive health system might just do that for us. If well crafted, this new system will exhibit responses that do not always require top-down directives. Rather, they can emerge from the collective interaction of AI and human. ~~The downside of emergent behavior is that it can be hard to understand the logic of unfolding events.~~

At this moment however, we are largely unequipped to either exploit or manage the cybersocial nature of our emerging smart and networked world. Our ambitions for the construction of a learning health system are still embryonic. Our understanding of the way blended AI-human collectives function is just as unevolved and requires interdisciplinary study[15]. There is a rich but not well integrated literature on the emergence of behaviour in social systems spanning sociology, economics, cognitive science, complex systems, network theory and AI [16]. The challenge for researchers is to harness this multidisciplinarity and to learn from~~study~~ the many natural cybersocial experiments already underway. For example, by modelling the effects of automated bots and opinion leaders in social networks, we should be in a position to explain emergent system behaviours such as vaccine refusal or the polarisation of political views [17,18]. Seeing cybersocial systems as "living" organisms reveals high level dynamic patterns of system growth, transformation, and even death [19].

When cognition is distributed, so is agency and accountability. We must ~~also~~ work to develop new strategies that can manage and exploit the unique characteristics of a cybersocial world. As we see with platforms such as Facebook, governing either individual or system behaviour is a challenging and moving target. Interventions that worked once can quickly become obsolete as new behaviours emerge, either from intentional workarounds, or as a natural evolution.

Strategies to create a system that behaves in the way we wish ~~can~~ could take a number of approaches (Table 1). The 'surprise' of emergent system behaviour is often related to the existence of unexpected connections and the way ~~such pathways~~they allow new interactions and outcomes. Surprises may take the form of novel behaviour, or interestingly, in the form of inertia to planned change[20]. The more hidden or 'dark' complexity there is in a system, the less likely we are to be able to predict its behaviour. ~~A systems engineering strategy to minimising surprise and maximising intended outcomes would first model the network of interactions that govern a particular system property, and then seek to reinforce critical pathways, or remove harmful ones.~~ Whenever new local health system interventions are attempted, it will be critical to not just assess their proximate impact, but also to ~~explore~~ model the connections of the local intervention to the broader system ~~whether they might have~~and identify unexpected ~~whole of~~ system consequences.

The research agenda to understand and exploit cybersocial systems presents a formidable challenge to healthcare as we do not routinely study our system from the perspective of complex networks, we do not think about the behavior of human-machine coalitions, and we are far more comfortable with targeted ~~and non-systemic~~ health service interventions. If we get this right, we will witness the birth of new working partnerships between humans and AI, new social dynamics, and new pathways to system change. The opportunity for healthcare is to conceive of a vastly different way to share and shape knowledge at scale, and to craft new, more connected ways of delivering care and living our lives. While our new and distributed cognitive health system ~~may~~ might not think in the way we traditionally understand it, it will surely have a life of its own[21].

Table 1: Cybersocial challenges and potential approaches to their governance in healthcare.

Characteristic	Implication	Management strategy
Dark Complexity: Much system complexity, originating out of interactions between humans and machines is unobserved. Whether because it is unmonitored, unmapped, or unmeasurable, hidden ~~complexity does~~complexity does not reveal itself until specific circumstances arise.	Responses to system interventions are inherently unpredictable. Definitive mapping of system components and all their interactions is not possible.	Localized mapping of safety critical system elements is essential, and testing of interventions in a wide variety of contexts and scenarios should minimize the risk of surprise. When new system elements are constructed, controlling the points at which they interact within the system will minimize unexpected interactions.
Redundant information flows: Complex interconnection makes it is hard to control information flows. Disintermediation (the removal of "middle men") is always possible where alternate information flows can be constructed around those who 'control' information flow.	Trying to control system behaviours by controlling information flow is easily outmanoeuvred. There are always opportunities for informational attacks such as message distortion or information overload.	Redundancy in information strategies decreases risks of being side-lined or attacked. "Innoculation" strategies may~~o~~ improve the resilience of decision makers to informational attacks. Distributed decision making by consensus may minimise the impact of localised attacks. Systems can be designed to help cope with unwanted or malicious information surges.
Rapid adaptation: Embedding distributed AI and human decision making into a complex adaptive system has the potential to make system adaptation unfold over a short timescale.	Early responses to system level shocks may minimize their impact and improve system resilience. Rapid change in best practice recommendations may improve the quality of clinical outcomes. Changes in best practice may lead to rapid and expensive obsolescence of services, technologies and therapies. Unwarranted adaptation may lead to waste and cause harm.	Close coupling of health service adaptation to real world outcomes, similar to post-market surveillance of medications, should help to minimise unwarranted adaptation or unexpected problems.

Emergent decision making: The actions of individual human and machine agents to satisfy local goals may collectively have unexpected impacts globally e.g. stock market crashes from high speed micro-trading.	Local behaviours can have global implications. Changes in the behavior of an AI external to but connected to your local system may result in changes to local decision-making.	Local actors need to be aware of potential global effects and develop strategies that manage these. This includes monitoring external dependencies and vulnerabilities, and unexpected changes in local decisions. Cognitive engineering, the shaping of network dependencies for decision-making may increase transparency and resilience.
Uncontrolled increase in complexity: Without ~~network~~ system-wide governance, there is no inherent mechanism to reduce complexity.	System characteristics such as inertia to change become more pronounced with time.	System 'apoptosis' mechanisms relying on AI can be envisaged from local (delete a guideline that is no longer accurate) to the global (decommission infrastructure or services that are not state of art) to create a 'forgetting health system'[22].
A society of systems: The health system sits alongside and interacts with many other systems including transport, education, finance etc.	Goals of different systems may conflict, neutralizing intended outcomes in one system, and triggering unwanted outcomes in others.	Explicit engineering of interfaces between critical systems, and explicit planning for managing goal conflicts between systems.

~~*Table 1: Some characteristics of cybersocial systems and potential approaches to their governance in healthcare.*~~

References

1. Turing AM. Computing machinery and intelligence. Mind 1950:433-60.
2. Coiera E. The fate of medicine in the time of AI. The Lancet 2018;392:2331-2.
3. Coiera E. Four rules for the reinvention of health care. British Medical Journal 2004;328:1197-9.
4. Hollan J, Hutchins E, Kirsh D. Distributed cognition: toward a new foundation for human-computer interaction research. ACM Transactions on Computer-Human Interaction (TOCHI) 2000;7:174-96.
5. Furniss D, Garfield S, Husson F, Blandford A, Franklin B. Distributed Cognition: Understanding Complex Sociotechnical Informatics. Studies in health technology and informatics 2019;263:75-86.
6. Suchman L. Human-Machine Reconfigurations: Plans and Situated Actions Boston MA: Cambridge University Press; 2006.
7. Stefanelli M. The socio-organizational age of artificial intelligence in medicine. Artificial intelligence in medicine 2001;23:25-47.
8. Wang F-Y. The emergence of intelligent enterprises: From CPS to CPSS. IEEE Intelligent Systems 2010;25:85-8.
9. Friedman C, Rubin J, Brown J, et al. Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. Journal of the American Medical Informatics Association 2014;22:43-50.
10. Friedman CP, Wong AK, Blumenthal D. Achieving a Nationwide Learning Health System. Science Translational Medicine 2010;2:57cm29-57cm29.
11. Dunn AG, Leask J, Zhou X, Mandl KD, Coiera E. Associations between exposure to and expression of negative opinions about human papillomavirus vaccines on social media: an observational study. Journal of medical Internet research 2015;17:e144.
12. Broniatowski DA, Jamison AM, Qi S, et al. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. American journal of public health 2018;108:1378-84.
13. Fraser H, Coiera E, Wong D. Safety of patient-facing digital symptom checkers. The Lancet 2018;392:2263-4.
14. Friedman C, Rubin J, Brown J, et al. Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. Journal of the American Medical Informatics Association : JAMIA 2015;22:43-50.
15. Rahwan I, Cebrian M, Obradovich N, et al. Machine behaviour. Nature 2019;568:477.
16. Gilbert N. Emergence in social simulation. Artificial societies: Routledge; 2006:134-43.
17. Bond RM, Fariss CJ, Jones JJ, et al. A 61-million-person experiment in social influence and political mobilization. Nature 2012;489:295-8.
18. Dunn AG, Surian D, Leask J, Dey A, Mandl KD, Coiera E. Mapping information exposure on social media to explain differences in HPV vaccine coverage in the United States. Vaccine 2017;35:3033-40.
19. Zheng M, Cao Z, Vorobyeva Y, Manrique P, Song C, Johnson NF. Multiscale dynamical network mechanisms underlying aging of an online organism from birth to death. Scientific Reports 2018;8:3552.
20. Coiera E. Why system inertia makes health reform so difficult. British Medical Journal 2011;342.

21. Barabási A-L. Linked: The new science of networks. AAPT; 2003.
22. Coiera E. The forgetting health system. Learning Health Systems 2017;1:e10023.

***Reply to Reviewers Comments**

I very much appreciate the input from editors and reviewers, who really have engaged with the piece and given me some clear and consistent feedback.

My responses to suggestions are detailed below.

Editor comments:

Editors (like the reviewers) wanted something more concrete. Cybersocial needs to be clearly defined …

> I now include a definition of cybersocial systems, based on previous usage of the term in the learning health system literature, and also sharpen up the discussion of distributed cognition which precedes it.

and clear clinical examples need to be given to illustrate the points made. What clinical advances can be expected and what clinical dangers may lurk.

> There are two changes I have made which should help:
> 1: I have slightly restructured the section headings and flow, as suggested by a reviewer, to improve the flow of the paper;
> 2: There is now a new box and paragraph which provides a description of what is unique about cybersocial systems. This should allow the health examples that follow to now be clearer.

Reviewer #1:

The paper presents a sweeping view of a system that will evolve if individual use cases of AI are implemented in an interconnected, and hopefully, thoughtfully-designed fashion within healthcare. It is very well-written and provides a refreshing take on what is becoming stale conversation around medical artificial intelligence (AI). I have a few suggestions to make.

1. A definition of cognition would be useful because there seems to be some threshold of technology and scale where a system starts to display facets of cognition as commonly understood.

> *Thank you, this aspect does require clarification. I think the critical idea needing definition is that of distributed cognition – the concept that for any system that includes cognitive agents, the effect of cognition emerges out of the interaction of environment, tools and individual cognitive agents. It is not that a distributed cognitive system is itself 'thinking'. I hope it is now clear in the paper hat when we think of a cognitive health system, it is in the sense of being a distributed cognitive system, not an individually cognitive one. Along with the cybersocial definition, I have rewritten the para on distributed cognition.*

The paper implies that a cognitive system differs from a learning system in that it can behave unpredictably or non-transparently, but such is a general property of any complex system especially as a consequence of scale. (In the same token, the table lists properties of any complex system, that is not specific at all to AI.)

> *Thank you, that is very useful. I have included a short new paragraph and accompanying Box to address the specific attributes that arise when we add AI into a distributed cognitive system. This also helps later on in providing a clear set of concepts that recur through the health examples, and when there is a discussion of the learning health system – which is a cybersocial system.*

The Wikipedia example does not help much because there is no AI at work in Wikipedia and its cognitive capacities were not detailed.

For space reasons I have removed the Wikipedia example. Interestingly AI is now being trialled in Wikipedia to help page editors identify unwanted page edits.

2. Does this particular cybersocial system emerge from an existing social system? If it does, it will be important to learn the new characteristics and phenomena around the existing social system that becomes cybersocial (vs. learning the behavior of a new system whose components would need to be defined). The Stanford Cyber Initiative describes a cybersocial system as cyber technology that interacts with an existing social system. Technically speaking, cyber systems such as computing systems, information systems and communication systems, are physical entities with specific elements for input, feedback and output. On the other hand, a social system is conceptual - a network of individuals that form an organization through norms, interactions and behavior. Given that the cybersocial system presented in the paper is a "distributed cognitive system that is characterized by a collaborative network of human and AI", what is its relationship with existing social systems?

A social system in which humans use technology like information and communication systems is usually described as a socio-technical system in the health informatics literature. A cybersocial system as described in this paper is a new kind of socio-technical system where some of the agents in the social system are AIs.

Does the medical AI cybersocial system specifically refer to
a. Emergent human behavior caused by human-driven AI? e.g. bots on Twitter driven by bad-actors pushing the anti-vaccination debate, and amplifying the voice of a bad actor
b. Unintended consequences of algorithms resulting in human tribalistic behavior? e.g. many people sharing anti-vaccination content, algorithms recognizing this and publicizing the content further, lending it an air of credibility that becomes self-reinforcing, and ultimately generating a movement
c. Algorithms independently operating and making opaque decisions for humans? e.g. AI-driven hospital/clinic/homecare triage algorithms
or all of the above?
This is not an exhaustive list, and each bullet point has a different aetiology and set of consequences, and deserves separate and robust discussion. Aggregating them all together under one umbrella might be confusing. The author might want to consider dividing the topic into a series, with multi-disciplinary involvement to explore each example to more depth.

Thank you. These are indeed some different possible behaviours or outcomes of a cybersocial system, and I agree others are possible too. I hope the addition of the new section and Box on the specific attributes of an AI enabled cybersocial system provides readers with enough structure to help them unpack the health examples that follow. The Rahwan et al. paper in Nature is a good place to look for more of this detail.

Given how much more could be said, as the reviewer suggests, I welcome the suggestion to consider writing something longer, perhaps a series, and agree that this would need to be multidisciplinary in authorship. Hopefully the current short paper can serve as a map for such a series, and future research papers.

3. As output of cybersocial systems, data may have questionable representation of social systems. The author states that "long term planning of urban geography, transport and energy, and the volume and type of health services in a community can all be informed dynamically by the thousands of digital behavioral experiments occurring across the population." This is a utopian dream given that a digital divide exists, and a cybersocial system in healthcare might further marginalize vulnerable populations. There should at least be some mention of this risk without going into an in-depth discussion of social justice in cybersocial systems.

Thank you. I add in a sentence as suggested to emphasise this point.

4. There seems to be some conflation of human-directed and human-independent phenomena,

especially when algorithms are presented as having independent agency (which ignores the human direction behind the scenes). The assertion of almost free-thinking and evolving trends in a traditionally unwieldy guild such as medicine and without oversight and evaluation needs to be qualified. The idea of "faster shifts to new best practices, enabled by machine learning and computational discovery systems, will also mean faster obsolescence of past practices" - is far-fetched in medicine. It trivializes the incredible hurdles required to deploy new technology even within most modern healthcare systems today, and minimizes the search for balance.

I appreciate this perspective. I agree that the traditional translational pathway from discovery through clinical trial and eventually to adoption is long and fraught. It is for this reason that proponents for a learning health system advocate the use of machine learning on observational data as an alternate to that pathway. It is not that 'machines' will tell us what to do, it is that AI-enabled analyses of observational data have the potential to produce new evidence at a volume and speed the traditional pathway cannot.

I add in a sentence to emphasise this, before exploring what I think are real and as yet undiscussed risks of speeding up best practice change.

5. The paper decried "message sharing over message security" as responsible for the anti-vaccination movement, but advocates for "community acquired data such as a person's location, movement and online engagements" to "help both in tracking disease, monitoring emerging population behavior and beliefs, and shape health services responses". Ethical issues regarding privacy and consent, and the need for transparency should at the very least be mentioned.

I have added a mention of privacy and consent issues in respect of the community tracking suggestions.

6. Finally, the paper might want to point to a large body of literature on emergent behavior within social systems in the fields of sociology and economics, and human factors literature on human-machine interfaces.

Thank you. I agree the pedigree for all of this is long and comes from many different disciplines. I now add an explicit sentence to reinforce that. I've also added a reference to Lucy Suchman's seminal work on situated cognition, the '06 Gilbert paper which tries to bring sociology and AI together, and already have a reference to the Rahwan et al paper in Nature which is encyclopaedic in its literature review.

Reviewer #2:

This is an interesting viewpoint in the context of AI. There is a number of strong messages in this piece to me the final paragraph says it all, whilst the meaning of some sections are less clear. I have the following comments:

1. In section two 'In complex environments … or to other humans'. This final part of the sentence seems not right (depending on the focus) or is incomplete and should be further explained in the manuscript.

I have rewritten and clarified the material in this second paragraph, with a more explicit definition of distributed cognition.

2. In the next section the author refers to 'a collaborative network of humans and AI'. This needs further attention as this is likely to be a crucial factor. Suggest to add items such as multilevel, specialisation/education to further elaborate on this.

I have rewritten the section by more clearly defining cybersocial systems, and as discussed earlier have a new section explicitly dealing with what is unique cybersocial *systems. The new paragraph explicitly looks at human-machine coalitions and some of their attributes.*

3. The description of the 'cybersocial' system lacks the human aspects such as empathy and emotional support, which are considered crucial in the health care. Even in the system without a focus on AI, these aspects are often lacking and as described now will further disappear in the cybersocial system. Please elaborate on that.

Thank you. I very much appreciate the issues raised here but did not feel there was room enough in the paper to elaborate, given the word count limit for this type of paper.

4. The section on opportunities and risks is at certain areas a bit vague. For example: "…inhabit a system which behaves unpredictably or non-transparently'. If this is the case (mainly referring to the unpredictably), I guess you will agree that these developments are not on the right track. Interested to hear your comments.

I hope the addition of the new material is much crisper and resolves this specific issue.

5. Again in the next section it is stated that '…even the creators do not have the ability to fully control or predict how it will behave'. It is of significant importance to avoid such issues in health care. A health care team has the ability to discuss certain actions/directions with each other and there will be a main care provider taking the final decision. If AI 'behavior' cannot be controlled, it is potentially harmful and dangerous in healthcare. These statements have to be revised or further explained in the manuscript.

The examples in this paper all happen at the population level, and there is no example suggested where a clinical team is somehow subverted in their care of a patient by a cybersocial system. There should be no inference in this paper drawn that we are creating 'out of control' AIs that will harm patients – rather we are creating a system whose behaviours at a macro level are not going to be easily managed.

6. Final paragraph of the 'opportunities and risk' section. Whilst the author refers to the combination of human and machine in the cybersocial system, there is no referral to the human aspect in this paragraph (specifically referring to the issue of flaws in the algorithmic advice'. The point is that although AI is provided to patients, there should always be a human aspect involved in that as well, in terms of education, structured follow-up and advice when to seek help/visit the hospital, despite the advice of the AI. This aspect is missing in the paragraph and suggest to add.

I agree that when a patient uses a technology like a symptom checker they should be embedded in a human care system to ensure poor outcomes do not occur through error or misuse. However that ideal is not the case in practice today nor in emerging digital first services. The example reflects that reality.

7. In general, the true meaning of the Cognitive Health System, is not clearly described. Provide a sentence in the final paragraph where this System becomes clear and where the meaning of this manuscript should all come together. Also the author refers to the fact that 'the system not always requires top-down directives. In terms of safety aspects, I can understand the top-down approach. But what about patient-focussed AI and used by patients, there should be bottom-up approach in AI as well. Do you agree? I would suggest to add this in the manuscript.

I hope the clarified material in the first section now makes the meaning of a cognitive health system much clearer. I do agree that 'bottom up' is an alternative, and retain in this version, the phrase "collective action" to cover alternatives to top down.

8. Final paragraph: I suggest to add the following to the second last sentence: "… more connected ways of delivering care and continue living healthy lives." This relates to increased

uses of technology such as Facebook, which the author refers to, and at the same time creating more solitude and some kind of a 'fake world' The latter issue is not well addressed in the manuscript and should be incorporated.

Thank you for this suggestion. As reviewer 1 indicate there is a huge territory to cover for this paper, and I certainly agree that social media may have negative impacts through social isolation and disengagement with the physical world. I think this is a substantive issue, independent of the specific point of this paper, and would require more than a few sentences to effectively summarise the growing literature in this area.

Reviewer 3:

There are a lot of really good ideas in this piece, but the presentation is disorganized. This manuscript may have significant contributions if the author can revise to clearly articulate the key messages in a methodical manner.

It seems the main message is to convey that the emerging cybersocial learning health systems of the future (or present) are quite different from existing learning health systems or social systems, and that there are important challenges and opportunities. In the current version, the definitions, differences, challenges, and opportunities are scattered throughout the manuscript rather than presented in a systematic way that builds an argument. I would suggest organizing as follows (order can be chosen for

(1) Begin with clear definitions of terms / systems described (and follow with consistent use of these terms) and how the cybersocial learning health system is/will be different from current learning health systems and social systems, as well as how current technologies fit into this framework. Use concrete examples of differences or features.

I believe the previously discussed changes to definitions and the new material should address this request.

Consider addressing the "natural history" of how the world will evolve from existing systems to future ones after clear definitions. There is a lot of discussion of current social media without a clear explanation of how those examples relate to the cybersocial learning system or artificial intelligence tools that are the focus of the paper. Make these relationships more clear.

The new material should now make that line of argument much clearer.

(2) Discuss opportunities.

(3) Discuss challenges.

(4) Discuss management strategies/recommendations.

(5) Summarize key points.

I have changed headings to better reflect this suggested structure.

I lot of the content is there, but the piece is confusing because the ideas are not presented systematically.