Beyond Silos: Why AI Regulation calls for an Interdisciplinary Approach

Sanjay Modgil (sanjay.modgil@kcl.ac.uk)
Reader in Artificial Intelligence, Department of Informatics, King's College London
Visiting Professor of Philosophy, Department of Philosophy, University of Milan

The lid has been prised loose from Pandora's Box, affording us a glimpse of the coming age of *Artificial Intelligence* (AI); an age that will dwarf the transformative impact of earlier technological revolutions. The potential benefits are enormous. From radically improved healthcare to revolutionising productivity in the workplace and enabling the green energy transition, AI is set to radically change our experience of the world.

But there are legitimate concerns about AI risks. Indeed, researchers, developers and business leaders alike, have called for more focus on regulating AI so as to mitigate against long term existential threats. These include the *indirect* role of AI in amplifying *existing* existential threats. For example, AI's role in subverting concepts (such as the very idea of 'truth' and 'facts'); concepts that underpin social consensus and a shared understanding of the world around us. And if we cannot agree on the facts, then what hope is there for collective action to address global threats such as climate change. Indeed, the recent startling advances in the capabilities of large language models (LLMs) such as ChatGPT, that almost no one expected to see in so short a time scale, has prompted calls for pausing further development of large scale AI systems. However, existing regulation is ill-equipped to deal with the unconstrained use of LLMs in generating misinformation that will then massively pollute our already contaminated informational ecology. And the prospects for *reactive* regulation are slim, given the widespread availability and uses of LLMs, and the pressures on companies to not fall behind in the race to profit from their commercial potential.

On the other hand, there are many who claim that warnings about long-term existential threats are overstated, that we are taking are eye of more immediate threats, and that unwarranted fear-mongering may result in over-regulation that then hampers AI innovation and the benefits that AI will bring. This more sceptical view, in tandem with worries that local restrictive regulation will handicap commercial exploitation and competitive advantage, is steering regulation to a more light-touch destination.

But should existing regulatory proposals, such as those currently proposed by the UK and EU, err on the side of being light-touch? Perhaps a more nuanced understanding of how AI may *amplify* existential risks can help answer this question and inform a more imaginative approach to AI regulation. An approach that, in contrast to these current proposals, is centred around an *inter-disciplinary* advisory group – a *SAGE AI* if you will – that in addition to AI researchers and technologists, includes anthropologists, philosophers, psychologists, social and cognitive scientists, economists, representatives from civil society e.t.c. The role of such a SAGE AI would be to promote and monitor ongoing interdisciplinary research into the short, medium and long-term societal impact of AI. Such a body would review and consolidate this research to advise regulatory authorities, while continually engaging with AI researchers, developers and businesses. The

hope would be that a SAGE-AI helps shape AI regulation to anticipate its development and uses, *before* AI systems are launched and made widely available.

Consider the significant societal challenges we are *already* facing: the polarisation of societies into rival "tribes" with increasingly entrenched political and cultural beliefs. Could regulation, informed by a SAGE-AI, have helped mitigate the role that social media's use of AI filtering and recommendation algorithms has played in exacerbating our contemporary post-truth polarised predicament? To answer this question, consider the following interdisciplinary understanding of how these algorithms effectively operationalise theconfirmation bias– the human instinct to selectively attend to evidence and opinion that support, and so further entrenches our beliefs.

Our distant ancestors lived in small farming communities, in which the confirmation bias might have served to entrench these groups' shared *tribal beliefs*; that is, beliefs relating to values, governance, resource allocation, religion, mythology e.t.c. The effect would be to strengthen bonds amongst tribal members, and so promote cooperation and a shared resolve to repel incursions from rival groups. We have thus evolved to experience dopamine mediated rewarding feelings when our tribal beliefs are confirmed.

Our ancestors relied only on each other to mutually reinforce tribal beliefs. However, with the internet, the available information is now not only vastly greater, but has the potential to expose us to misinformation and extremist views on an unparalleled scale. Moreover, the "attention economics" of the internet, and in particular social media platforms such as Facebook, has incentivised how this vast repository of online information is filtered for our consumption. Our search and click histories effectively provide a profile of the tribal beliefs that we engage with. Algorithms then selectively feed us with more of the same, and the rewarding feelings accompanying confirmation and reinforcement of our cherished tribal beliefs entices us to spend more time online, increasing exposure to revenue generating adverts.Thus, AI filtering and recommendation algorithms are technological incarnations of our innate confirmation bias, selectively feeding, confirming and entrenching our existing opinions. In concert with the increasing amounts of online fake news and misinformation, and a host of other societal developments, these algorithms may then contribute to leading us down rabbit holes to ever more extreme versions of these beliefs, and polarising societies to the detriment of societal well-being[1].

But why does this technological outsourcing of our confirmation bias, massively amplified by the ubiquitous availability of vast amounts of reinforcing information, no longer serve our interests as it once might have done? Because groups with shared tribal beliefs, such as anti-vaxxers and climate change deniers, no longer live side by side. Instead, rival tribe members live in the same neighbourhood, but with increasingly rigid and divergent understandings of their shared habitat. And unflinching disagreement about the facts undermines consensus on practical solutions to known existential threats such as pandemics and climate change. (Are we confident that there would be sufficient vaccine uptake to guarantee herd immunity when faced with a far more virulent pandemic?).And in the absence of effective regulation, the available content for reinforcing and radicalising tribal beliefs is set to not only increase by orders of magnitude, but also undergo further massive pollution, given the widespread availability of LLMs for generating unlimited amounts of misinformation and fake content that is then posted online. An information apocalypse is nigh!

[1]The role of filtering algorithms in societal polarisation should not be over-emphasised. It is arguably one amongst a complex interplay of other societal factors, as exemplified by the Covid Pandemic.

We've seen how a relatively unsophisticated use of AI has, in combination with other societal developments, led to *unexpected*consequences. After all, the early internet pioneers harboured utopian dreams of aninformation superhighway that would*erode*barriers to a shared global vision, and not strengthen them! And as AI becomes more intelligent, and as we delegate more to AI so that it acts with more independence, we may reap other unexpected consequences. And as AI becomes more integrated into human society, these consequences will become more difficult to control and undo (consider our current limited options for addressing the socially divisive impact of social media and the free-for-all use of LLMs).But our ability to anticipate the long-term effects of a multi-faceted issue like AI's impact on society, to "expect the unexpected", will require an interdisciplinary SAGE AI shaping of regulation that strikes a balance between high-risk light-touch and overly restrictive heavy-handed approaches.

Contrast this proposal with what is currently on offer. The recent UK white paper on AI regulation proposes the empowerment ofexistingsiloed regulators to come up with tailored approaches to regulation for specific sectors. The EU AI Act proposes prescriptive legislation, spanning sectors and focussing on existing "prohibited" and "high-risk" AI systems. But neither have centralised mechanisms for feeding through analysis of AI's societal impact to those involved in regulation or working on AI R&D.That said, the recent UK AI safety summit did schedule multidisciplinary discussion around the societal impacts of AI. It remains to see whether the summit will instigate a redrafting of the regulatory landscape[2].

Looking to the future, what other issues could a SAGE-AI advise on? Philosophers, psychologists and researchers in Digital Humanities are now raising concerns about AI systems triggering the human 'anthropomorphic' instinct to ascribe rich human-like mental lives to entities, and in particular robots, that exhibit human like behaviours.

For example, companies now provide online AI romantic partners, and we are witnessing unhealthy attachments being formed with these virtual partners. But as we design humanoid robots that not only behave and speak like humans, but are human like in appearance, our anthropomorphic instincts will be very hard to resist. And in the sex industry, there will be an overwhelming commercial imperative to design sex robots that simulate arousal and reciprocal attraction, and so ratchet up anthropomorphic ascriptions of human-like feelings and emotions. But how will engagement with these humanoid robots, who may well be treated as subservient 'sex slaves' by their human 'owners', while simultaneously being thought of as conscious entities, affect the capacities of their owners to develop consensual and respectful sexual relations with other humans? After all, we are already witnessing the detrimental effects of overuse of internet porn on emotional and sexual relations amongst humans.

On the other hand, could robot use to support human carers (e.g., in the under-resourced care sector) conceivably be more nourishing, more acceptable by those who are being cared for, if their humanoid designs suggest that they genuinely care and experience empathy?But then how would this impact the extent to which we humans feel we are responsible for caring for our elderly? A SAGE-AI should be promoting and synthesising research *now*, into the impacts of anthropomorphism on human society, so as to address these kinds of questions.

Pandora's box is open and ChatGPT4 and other powerful AI systems have been released. There is no going back. We must minimise the risk of unexpected consequences, and forearm ourselves against those that we anticipate, while also positioning ourselves to reap the transformative benefits of AI. To do that we need smart regulation. An interdisciplinary understanding

[2]The first progress report from the UK government's Frontier AI taskforce is proposing an expert advisory body primarily consisting almost exclusively of technologists.

of AI and its impact on society needs to be front and centre when it comes to thinking about AI safety and regulation.