SCAI
SINGAPORE CONFERENCE ON AI

Preliminary Conversations
Towards AI for the Global Good

THE SCAI QUESTIONS

CONTENTS

ABOUT SCAI

AI has the potential to enhance our quality of life, revolutionise industries, and transform the way we live and work. However, there are many potential challenges that may constrain our ability to harness the technology to benefit societies and people, such as accuracy, bias, and resource-efficiency.

To overcome these challenges, the Singapore Ministry of Communications and Information and Smart Nation Group, in partnership with the Topos Institute, organised the inaugural Singapore Conference on AI for the Global Good, or SCAI, from 4 to 6 December 2023. The conference brought together 42 distinguished experts from various fields of academia, industry, and government.

Drawing on their diverse domains of expertise, delegates explored and articulated critical questions of AI that, if answered, will enable the development and deployment of AI for societies to flourish.

FOREWORD TO

THE SCAI QUESTIONS

These Questions were conceptualised and written by the SCAI delegates over the 3 days of the conference, using a process designed to synthesise diverse views from experts. Each of the 12 SCAI Questions is envisioned to be a comprehensive articulation of a foundational, yet tractable area of AI development and/or deployment. The 12 SCAI Questions taken as a whole, are meant to be a holistic formulation of the challenges that should be addressed by the global AI community to allow humanity to flourish.

Good questions are hard to frame, especially in a domain as emergent and boundary-spanning as AI. The SCAI Questions are the best-effort of the delegates gathering together, debating and then consolidating their views over 3 days in Singapore. They are certainly not final, and we invite commentators and researchers to use these Questions as a springboard for further research, collaboration and innovation.

In terms of format, each Question begins by stating upfront the context and assumptions which the delegates had in mind, followed by an elaboration of the possible approaches to answering the question, known challenges, and ways we might recognise progress.

The SCAI Questions are a collective and collaborative product of the conversations amongst SCAI delegates. They do not necessarily represent the views of individual participants, or the organising parties of SCAI.

The ordering of the SCAI Questions is solely for ease of reference and does not reflect a hierarchy of importance.

For referencing or citing this document in academic or professional contexts, please use the following format:

Singapore Ministry of Communications and Information & Smart Nation Group, in partnership with Topos Institute. (2023). *Preliminary Conversations Towards AI for the Global Good: The SCAI Questions*. Proceedings of the Singapore Conference on AI for the Global Good, 4-6 December 2023, Singapore. Available at: https://www.scai.gov.sg/findings.

SCAI QUESTION 1

SCAI QUESTION 1

RELIABILITY & TRUSTWORTHINESS

How do we ensure that AI models and systems are reliable and trustworthy?

Context & Assumptions

AI systems are increasingly being used for decision-making across various fields, including critical, high stakes areas like medicine. However, current systems are not always reliable nor trustworthy. For instance, language models often "hallucinate", producing outputs that are inconsistent and not grounded in reality. The result is that human users cannot trust the output and cannot rely solely on the models to make decisions.

For an AI system to be reliable, it should consistently produce outputs that align with a specific, well-defined set of requirements, even when deployed in new or changing environments. For example, a self-driving car should adhere to a specified performance standard in terms of speed and stability under different weather and traffic conditions; and a medical language model should provide accurate and up-to-date medical information.

Trustworthiness, on the other hand, is a broader function of the relationship between the AI system and its human users. For example, it includes whether the behaviour of an AI system is consistent with its users' expectations; whether it is in accord with human norms of fairness and ethics; and if it is robust to adversarial conditions. For instance, a self-driving car might be

deemed untrustworthy if it drives unlike a human, or a language model might lose trust if it is found to be biased against certain demographics.

Question

How do we ensure that AI models and systems are reliable and trustworthy?

Evaluation is a key aspect of reliability. How do we design specifications for complex, open-ended tasks, and then evaluate models against them? Ideally, these specifications should be standardised, reproducible, lightweight, but also as close as possible to actual downstream tasks; they should capture all relevant aspects of the desired model behaviour (e.g., not just average accuracy but also notions of bias); and they should be evaluated in environments that are reflective of real-world settings, including end-to-end tests with users as appropriate. Designing such specifications is particularly challenging for generative AI systems, where the output space is large and automated evaluation is difficult. Beyond the standard approach of empirical testing, an open question is whether we can provably verify the reliability of white-box models, which is especially relevant to the development of defences against adversarial actors.

To achieve trustworthiness, reliability is necessary but not sufficient; we also need to consider the interaction between the full system and the user. An open question is how to design systems so that the decision making process by the system is observable and interpretable to the user. The structure of the interaction and the communication between the system and the user can significantly influence the level of trust users have in the system and must be designed appropriately. Finally, trust requires that the specification itself be correct and useful within the user's context.

Systems operate continuously – trust is engendered when systems maintain reliability in novel environments and remain current with evolving knowledge. Providing reliable performance over time depends on recognition of distribution shift, detection of out-of-distribution queries, and the ability to respond appropriately. An open question is whether this adaptation necessitates an understanding of the world, including context, causal structures, and implicit motivations. Issues of uncertainty, calibration, and security are also pertinent. How well an AI system can handle uncertainty and how accurately it is calibrated are crucial for its effective functioning. Security considerations, especially in the face of potential adversarial attacks, cannot be overlooked.

Indicators of Progress

A key indicator of progress will be the development of standardised benchmarks that, as elaborated above, capture the relevant aspects of reliability and trustworthiness in a broad range of real-world applications of AI, and in particular in open-ended tasks that require more sophisticated evaluation. As technical progress might overfit to existing benchmarks, established processes for continually creating more diverse and realistic benchmarks would be another indicator of progress.

Another indicator of progress will be a codified set of principles for "Design for Reliability," akin to "Design for Manufacturing", which will allow AI systems to be specifically designed with the goal of meeting specifications that go beyond statistical performance. These principles will include developing methods for uncertainty quantification and introspection, continual learning, out-of-distribution robustness, explainability, and enabling AI systems to recognize and communicate the limits of its knowledge. For trustworthiness in adversarial environments, the development of methods for provably verifying model outputs will be another indicator of progress.

Finally, a significant indicator of progress will be the deployment of AI across various domains, gradually increasing in scope, autonomy and operational duration. This would collectively show the advancement and maturity of AI systems in being both reliable and trustworthy.

SCAI QUESTION 2

SCAI QUESTION 2

DATA COLLECTION & SHARING

How can we create a data collection and sharing ecosystem that produces high-quality data for AI, which can be shared and exploited within and across countries?

Context & Assumptions

High-quality data leads to high-quality AI. Training large models currently requires large amounts of high-quality data. To have high-quality data, one needs to consider what would be appropriate data governance, technologies, and infrastructures to manage the challenges of data collection, deal with data fragmentation, and support data integration. In addition, there is a need to address issues concerning legal real time continuous data (e.g., sensor data), fact editing, and challenges around model collapse and reproducibility. Data about the construction and performance of AI models themselves is an important asset.

Building good datasets has been problematic. In some domains, acquiring data can be a challenge. For example, in healthcare, developing good models requires diverse, high-quality, accurate datasets from the local community, but this is difficult to achieve given the sensitivity and privacy of such data; these challenges persisted even during the COVID-19 pandemic, where data was essential to combatting its spread. Also, user generated content, a goldmine of potential insights, is often proprietary. Companies collecting this data are cautious about sharing, as it has

competitive value and comes with privacy concerns and legal obligations.

Despite the challenges, some datasets and resources are maintained to high standards. This ranges from collaborative maintained resources such as Wikipedia to cultural heritage data, e.g., Europeana, the European digital cultural platform, which allows museums, galleries, libraries, archives across Europe to share and reuse digitised, standardised cultural heritage images such as 3D models of historical sites, and high-quality scans of paintings. Additionally, biomedical data, such as genetic information, clinical trial results, and disease data, are typically well-catalogued and publicly available, fostering scientific collaboration and research.

Question

How can we create a data collection and sharing ecosystem that produces high-quality data for AI, which can be shared and exploited within and across countries?

A robust data collection and data-sharing ecosystem will also allow us to address the following key issues:

- **Measuring data quality**: Principles guiding the measurement of data quality are essential for fostering reliable AI models. Adhering to the FAIR principles (Findable, Accessible, Interoperable, Re-usable) lays the foundation for robust datasets. Additionally, for factors such as diversity, representativeness, openness, trustworthiness and safety, the incorporation of mechanisms for both prevention and assistance in the unlearning processes of AI models are crucial. These principles collectively contribute to the integrity and efficacy of AI models and applications.

- **Data valuation**: Relatedly, the assignment of value to data is important in any effort to enable effective sharing. Data valuation can help to guide the selection, management, and the stewarding of data.

- **Scaling data**: Large language models require massive amounts of text data for training. Other frontier models require significant data in other modalities; image, video, audio etc. Acquiring and curating such large and diverse datasets can be logistically challenging and resource-intensive. Scaling data to accommodate large volumes necessitates a federated approach. Discovering and linking data across domains, maintaining standards to enable data sharing, maintaining a balanced representation across various demographics, and exploring the generation and use of synthetic data are potential approaches.

- **Data privacy**: Ensuring data privacy is paramount in a data-sharing ecosystem. Implementing robust privacy measures involves anonymisation, encryption, and compliance with international privacy regulations. Striking a balance between data utility and privacy protection is essential for fostering trust among data contributors and users.

- **Data rights**: Establishing clear ownership frameworks ensures responsible data stewardship and facilitates ethical data sharing practices. Some of the primary challenges in managing data rights pertains to the constraint imposed by political boundaries—commonly known as data sovereignty. Additionally, navigating data copyrights requires understanding intellectual property laws and implementing mechanisms to protect creative elements within datasets.

- **Data reproducibility**: Data reproducibility is critical for the scientific community and AI practitioners. Implementing practices such as sharing code, documenting methodologies, and providing access to raw data enhances the reproducibility of AI research. Open and transparent practices will continue contributing to the credibility and reliability of AI models and findings.

Indicators of Progress

For data to be effectively exploited to build AI for good, we will need to consider a range of levers, including governance and technological approaches.

The governance approach, including setting up relevant regulations, incentives, subsidies, data formatting and sharing agreements, can encourage the creation and flow of data. For example, one can mandate that datasets that are created through the support of federal funds should be publicly available in a standardised format. Furthermore, one can encourage the adoption of an agreed methodology to describe the origin, nature, and use of data. The UK Biobank is one such example. Defining a taxonomy for data can also help classify data more effectively for appropriate privacy classification and data licensing.

In addition, there should be equivalent focus on novel technical approaches to achieve high-quality data for AI. This would involve research investments. Current examples include privacy enhancing/preserving technology and development of Trusted Research Environments (TRE). TREs allow for model development and deployment by leveraging data in a decentralised and secure manner. New technologies to enable data provenance can also enable safer development of AI, especially large foundation models, so as to allow for appropriate model unlearning, licensing, watermarking etc.

There should also be a focus on measuring the progress of enabling high-quality data sharing, which in itself is a non-trivial challenge. There will be cultural issues, liability concerns, political and jurisdictional boundaries that come into play that might impede creating and sharing data. There needs to be more recognition of the progress of such work, including the measurement of high quality data sharing and quantifying the friction of data flow, metrics of unlearning data, as well as data quality parameters such as data diversity and uniqueness, accessibility, etc.

SCAI QUESTION 3

SCAI QUESTION 3

GOVERNANCE STRUCTURE & REGULATORY MEASURES

What are optimal governance structures and regulatory measures for AI?

Context & Assumptions

Governance and regulation have a key role to play in shaping the direction of development and deployment across the spectrum of different AI technologies and applications. Governments should provide the conditions for confident AI innovation and adoption and for preventing AI-related harms. The purpose of public policy is to protect the public interest, and we assume that effective, efficient and legitimate governance and regulation is a precondition for trusted and sustainable advancement of AI, rather than representing an obstacle to innovation. At the same time, the multi-faceted, complex, and border-crossing nature of AI raises challenges for achieving effective, efficient, and legitimate governance structures and regulatory measures.

Success requires coherence and integration of governance structures and regulatory measures across multiple dimensions. These should ideally encompass different sectors and regulatory remits, various points of intervention across the AI lifecycle and AI value chains, different types of

regulatory tools, and several levels of governance (municipal, national, regional, and international/global).

In order to achieve this integration, and to do so with legitimacy, it is important that processes of designing governance structures and regulatory measures are inclusive of perspectives from the diverse groups in society whose interests are at stake and whose expertise can help identify solutions, including users, businesses, policymakers, civil society, and academia.

Legitimacy also depends on ensuring checks and balances, enforcement and accountability mechanisms, as well as the prevention of abuses of power by those who deploy and use AI.

Question

What are optimal governance structures and regulatory measures for AI?

Answering this question involves determining the appropriate combination of structures and measures along several dimensions:

- **Legitimacy**: Which parties have legitimacy to establish governance mechanisms in a given jurisdictional context (government vs. non-government actors; collective or representative organisations)?

- **Sectors and regulatory remits**: What is the optimal combination of governance and regulation designed to apply to AI horizontally vs. in specific sectors vs. in specific use cases/applications?

- **Types of governance tools**: What is the optimal combination of different regulatory levers, such as legislation, mandatory codes of practice, voluntary frameworks, standards and principles, public procurement rules, and other tools which can be used to achieve desired practices and outcomes?

- **Points of intervention**: What is the optimal combination of structures and measures that target different stages in the AI lifecycle, different segments in the AI value chain, and should these be implemented as ex ante requirements (e.g. as preconditions for licensing or regulatory approval) or as ex post actions (e.g. requirements to take remedial actions after the fact)?

- **Levels of governance**: What is the optimal combination of municipal, national, regional, and international/global measures? To what extent is international/global alignment or harmonisation desirable and achievable (as opposed to instances of divergence that are unavoidable due to fundamental differences in political ideologies and values, and respect for state sovereignty)?

- **Accountability**: What do enforcement and accountability look like? Who should be responsible for enforcement, and what mechanisms are appropriate (e.g. criminal sanctions, regulatory penalties, access blocks and bans)?

Indicators of Progress

There are some significant challenges in determining the optimal governance or regulatory approach for AI development and use. These include a **lack of transparency and access to information** on how AI models are developed, governed, and deployed, especially in the private

sector; a **failure of coordination** between various agencies or departments in governments or organisations involved in exercising governance or regulatory functions or processes; a **lack of stakeholder inclusion** in developing governance and regulatory frameworks; a **lack of clear and effective reporting, enforcement, and accountability mechanisms; abuse of or misalignment with incentives**, whether at the political or commercial level; **insufficient resourcing, capabilities, and expertise** within governments or organisations in designing and implementing governance structures or regulatory measures, and in ensuring compliance; and **regulatory lags** given the ever-changing and fast-moving nature of AI.

While these are not insurmountable, careful attention must be paid to addressing and overcoming these challenges, in order to prevent them from becoming key stumbling blocks to progress in answering the question.

Possible strategies to make progress on this question include:

- Systematic mapping, gap analysis and guidance for relevant existing laws and legal principles, and governance structures and regulatory measures in individual and regional jurisdictions.

- Adopting new laws and regulations to fill the gaps that existing law does not cover.

- Establishing new forms of collaboration and coordination between regulatory bodies across different sectors, remits and jurisdictions.

- Systematic approaches to developing score cards and evaluation frameworks for governments and companies.

- Designing reporting, liability and accountability schemes.

- Advancing development of international standards that enable interoperability between regulatory requirements and governance frameworks established in different jurisdictions.

Measures of progress in answering the question include:

- Emergence of clear, effective, efficient, and legitimate governance frameworks for the use of AI, with appropriate scrutiny (including by the media), accountability, and checks and balances on the use of AI.

- Implementation of measures specifically aimed at protecting the public interest in the context of AI development and deployment, including product safety, citizen rights, and rules relating to the use of AI in providing public services.

- Whole-of-government and cohesive approaches in policymaking and regulation, involving all relevant government departments and regulatory agencies.

- Increased transparency of and access to information on how private sector models are developed and their governance and implementation frameworks, as well as on governmental development and use of AI.

- More countries adopting explicit and clear positions on AI governance and regulation, whether by explaining how existing law and regulations apply to AI, and/or by introducing new laws and regulations for AI.

- Increased, and new forms of, international collaboration, cooperation, and capacity-building on AI regulation, standards, and other governance frameworks, mechanisms and tools.

- Development and promulgation of international standards that can enable interoperability between frameworks and approaches in different jurisdictions.

SCAI QUESTION 4

SCAI QUESTION 4

SOLVING SCIENTIFIC PROBLEMS

How should we advance AI to solve scientific problems that are critical and beneficial to humanity as a whole?

Context and Assumptions

Throughout human history, significant advances have been made possible only by scientific progress. Scientific discoveries such as electricity, penicillin and semiconductors that have played key roles in progress have been driven by scientific discoveries.

Harnessing the power of AI systems offers a promising avenue for addressing some of society's most difficult problems that remain unsolved. These intractable problems which significantly impact societal and individual longevity, are ultimately solvable but have persisted for decades, such as those involving climate change and complex biological systems. AI holds the potential to bring innovative self-directed approaches that will assist humanity in fundamental ways. For example, in climate change, advances in carbon sequestration will enable humanity to deal with the impact of global warming due to fossil fuels, and in complex biological systems, advances in genome sequencing and editing could help us understand human biology and develop cures for disease.

There are, however, some challenges. There is a general and systemic lack of integration of foundation models and exploratory methods that generate and examine new hypotheses. There is also a need to develop logical and causal inference mechanisms in AI neural systems, as well as adequate funding for the specialised infrastructure required for AI development.

Question

How should we advance AI to solve scientific problems that are critical and beneficial to humanity as a whole?

In considering this question, we also need to think about how scientists and researchers across the world can come together to harness AI and prioritise resources in applying these powerful systems to scientific problems.

Indicators of Progress

In the short-term, a notable indicator of progress includes the growing evidence of increasing cross-disciplinary and cross-border collaboration and cooperation. Addressing traditional boundaries between different scientific and industrial areas can play a significant role in facilitating access to a diverse and extensive range of scientific and mathematical corpus. In the longer term, international research collaboration, supported by multilateral funding, has the potential to yield results that benefit the global community. Importantly, this would allow us to integrate scientific knowledge into AI models to increase their reliability and accuracy, while using less computation.

A key indicator of progress will be the emergence of theorem-solving AI systems that are applicable to a wide variety of scientific areas of interest, and allow us to draw on cross-disciplinary scientific knowledge at a scale previously unavailable but with the potential to transform scientific development. We look forward to a time when we are able to see high impact scientific papers being written by such AI systems with application in areas such as carbon sequestration, seasonal climate prediction, deciphering the human ageing process and new materials design.

SCAI QUESTION 5

SCAI QUESTION 5

MODELS & ARCHITECTURE DERIVED FROM NATURAL INTELLIGENCE

How do we leverage developmental models and architecture derived from natural intelligence to create new paradigms of AI?

Context & Assumptions

Human intelligence and cognition have capabilities and performance characteristics that are currently unmatched by the best AI systems. While some AI systems outperform specific human capabilities, they fall short in generalisation capabilities and learning efficiency. Natural intelligence is far more flexible, adaptive, responsive, and energy efficient. The brain and our understanding of its functional and cognitive architecture offers a possible reference to implement an intelligence with these performance characteristics.

The functional and developmental organisation of natural intelligence has a considerable impact on how intelligent capabilities form and perform. Hardware substrates in the brain (e.g., neurons) differ from the artificial hardware substrates (e.g., processing units). For instance, the

hierarchical structure of spatial reasoning in the brain, between cortical grid cells and hippocampal place cells, provides considerable computational efficiency and robustness that robots are only now beginning to match. In addition, the ability of the brain to perform in-memory computation provides significant efficiencies that cannot be matched by the current separation between computation and memory in existing CPUs and GPUs. There is also evidence that the cognitive architecture of the brain is genetically encoded at birth - the core knowledge hypothesis suggests a strong prior over concepts such as places, objects and motor skill which is already present in biology.

The increased understanding of the functional architecture and performance of the brain provided by cognitive neuroscience can give us new paradigms to develop more capable forms of artificial intelligence. The cognitive sciences (neuroscience, psychology, linguistics, philosophy of mind, anthropology and artificial intelligence) study different aspects of natural intelligence that can be used to inform the design of more capable AI systems. Conversely, progress in AI creates opportunities for understanding brain functions, human cognition, psychology, and development.

Question

How do we leverage developmental models and architectures derived from natural intelligence to create new paradigms of AI?

Answering this core question is tightly coupled to the following additional questions:

Evaluation questions

- Which aspects of natural intelligence cannot currently be replicated by existing AI approaches? This is a moving target, but it is crucial to understand precisely how natural intelligence outperforms existing artificial intelligent systems, and at which tasks.

Structural questions

- What is the right functional decomposition of intelligence that enables these levels of performance and capabilities? The functional relations implicitly define a structure that may be reflected in the structure of the brain.

- What are the intermediate hierarchical structures in the brain that organise neurons into functional reasoning and cognition? While cognition and intelligence do not need to be implemented by neurons, there are existing models of artificial intelligence represented using spiking neural models. An additional intermediate hierarchical structure is required to organise artificial spiking neural models into purposeful computation to allow program synthesis.

- The functional decomposition and cognitive architecture of natural intelligence imply specific and powered inductive biases. What are these inductive biases inherent in natural cognitive architectures?

- How can biological models of motor skills be acquired and composed by artificial intelligence? It is clear that natural intelligence can acquire low-level motor skills efficiently, and incorporate these skills as concepts into higher-level reasoning. There is further evidence from evolutionary biology that developing these low-level skills took considerably more evolutionary time than higher-level reasoning, and may be considered

the true cognitive substrate of intelligence.

Performance questions

- Can computational architectures inspired by the cognitive architectures of the brain change, adapt and evolve as easily as the brain does? The power consumed by in-silico intelligence dwarfs the power consumed by biological intelligence, but with a fraction of the performance.

- Can computational architectures inspired by the cognitive architectures of the brain match the energy efficiency of the brain? There is considerable evidence that when a biological agent encounters new scenarios, it is quickly able to adapt to the scenario by reusing previous experience.

- Do the cognitive architectures of the brain implicitly encode an inductive bias that is aligned with human values and judgement?

- How can we ensure and maintain alignment between artificially intelligent agents and humanity?

Indicators of Progress

We expect that the best approaches to answering these questions will include:

- Leveraging development models of natural intelligence and insights from neuroscience and cognitive science and implementing these models in architectures inspired by them. We expect that one candidate form of these models and architectures will be hierarchical compositional probabilistic models that can be reused.

- We also expect that one candidate form of these models that allows the structure of the architecture to be learned and to be adapted from, includes neurosymbolic representations.

- We will also require new theories of model integration (which is not the same as interoperability) and techniques for learning to be used for integrating component models.

- Designing neuromorphic AI software and hardware, and examining the benefits that can be gained by neuromorphic and neurosymbolic approaches.

We expect that the best approaches to measuring progress in answering these questions will be:

- Benchmarking brain-inspired architectures and cognitive systems on tasks relative to human/natural performance.

- Benchmarking on task specialisation, generalisation and few-shot learning.

- Benchmarking developmental models, that allow the cognitive architecture to develop and adapt over time. For example, tasks that have an internal hierarchy with different levels of abstraction that are currently hand-specified (e.g., perception systems driving task-and-motion planning systems) can be derived automatically using developmental models derived from natural intelligence.

SCAI QUESTION 6

SCAI QUESTION 6

VALUES & NORMS TO ALIGN AI: ELICITATION & IMPLEMENTATION

How do we elicit the values and norms to which we wish to align AI systems, and implement them?

Context & Assumptions

Increasingly capable AI systems are being used to perform more complex sequences of actions without human supervision. We collectively need to know how we want them to behave and how to ensure they do so. This has historically been described as "the alignment problem". However, the aim of aligning systems to "user intent" or to "human values" is a double-edged sword. Users might have malicious intents; humans can have abhorrent values. In addition, and not coincidentally, the project of AI alignment has been pursued in a narrowly technical way, without drawing enough on broader expertise (e.g., from the social sciences and humanities), even as other areas of responsible AI have done more to integrate their research with other fields. There is an urgent need to develop an agenda for AI alignment that draws on this broader understanding to ensure that AI systems behave appropriately.

Question

How do we elicit the values and norms to which we wish to align AI systems, and how do we implement those values and norms?

Indicators of Progress

Eliciting the values and norms to which we wish to align AI systems is not a novel problem. It is simply the challenge of reaching a collective decision on matters of common concern. The first stage is to provide the theoretical and empirical resources for public debate and individual decision-making:

- We need well-grounded research anticipating potential societal impacts of more capable AI systems. Social scientists and computer scientists should collaborate to explore different possible futures for AI, and to learn from the rich experience with previously deployed systems to anticipate likely risks of future systems.

- We need clear articulations of familiar normative considerations that those potential impacts raise. For example, most societies already have clear, albeit disputed, views on values like discrimination, accountability, and transparency. The goal then is to apply and refine those values for this particular application.

- We need a theoretical approach to unfamiliar normative considerations raised by those potential impacts. Some questions raised by more capable AI systems will not come with ready-made answers. For example, if A delegates an action to B, then do the reasons that

apply to B when B acts depend on whether B is a human or an AI agent? Which behaviours are acceptable or unacceptable from AI agents? When interacting with humans, one set of rules might apply, but which rules will apply in multi-agent situations? Or, can extremely capable AI systems ever be consistent with democratic government? More generally, should societies even be pursuing the goal of AI capability beyond a certain threshold?

The second step is to use our existing resources for collective decision-making and resolving moral disagreement. This means recognising at least three distinct layers of normative guidance, with different collectives being appropriate to decide on different layers. As with most other societal decision-making, this will involve some "constitutional" norms that are relatively settled, and others that should be regularly revisited and revised:

- Some minimal norms should be decided at the global level, in the same way as the global community decides on certain basic human rights. Which highly capable AI systems should nobody be able to produce? What are the very minimum expectations for the behaviour of AI systems, on which the whole world can decide?

- More substantive norms should be decided at the level of nation-states or other sub-global political units (e.g., the EU). By analogy, while all states in principle affirm the same basic human rights, they all have different approaches to civil and political rights. Operative questions: are there any AI systems that we want no one to produce? What are the minimum expectations for the behaviour of AI systems, on which we as a political community can decide?

- Remaining norms can be the object of individual, or (sub-state) collective choice, including by companies. Operative questions: given the constraints described in A and B, which AI systems do we want to produce? Which behaviour (not just minimum

> expectations) do we want to see in our AI systems? Analogy: virtue in a person. If all you ever did was violate nobody's human rights and not break the law, then that would not on its own speak well of you as a person. You also might aim to be honest, loyal, loving, conscientious, etc. Those who build AI systems that can act (in effect) autonomously should want to do more than the bare minimum.

Our means for resolving moral disagreement and making collective decisions are often compromised, and direct action or institutional innovation may be needed. AI itself may help unblock decision-making, for example, by supporting participatory or deliberative democratic processes (especially when deciding on values beyond basic human rights and legal compliance). But we must avoid using technology to replace politics instead of augmenting it.

How do we implement these values and norms? We recommend sociotechnical methods that complement technical methods in computer science with expertise from the social sciences and humanities.

All AI systems will be deployed by people in a social and political environment. "Aligning" this sociotechnical system can be achieved through interventions on each of these elements, e.g., placing models in more complex systems that mitigate some of their risks; training users to avoid automation bias; thinking about institutions: for example, can we (should we) reshape political institutions and AI systems so that AGI, if achieved, does not directly undermine democracy?

We also need pre-deployment sociotechnical evaluations that consider the harms caused in actual use, rather than in isolation from broader social systems. And we need to advance adversarial testing (red-teaming) beyond simply querying the model to elicit naughty text, developing instead complex multi-agent simulations that test for dangerous capabilities. This may entail some "gain of function" AI research, which may require methodological innovation.

Ultimately, however, we want to produce systems that are designed to behave appropriately (given the three stages of norms described above), and can be counted on to do so (preferably provably). Designing AI systems that will implement these values directly is therefore essential. Collaboration between computer scientists and other fields will provide fresh perspectives on alignment methods, and suggest new research directions.

Any method that applies a thin fine-tuning adjustment over the top of a pretrained model is unlikely to be robust to adversarial attacks of different kinds. Also, in learning from human feedback (e.g., Reinforcement Learning for Human Feedback (RLHF)) for language models, the behaviour being evaluated is identical to the behaviour being shaped; but if LLMs are used as the executive control centre for more complex systems (i.e., agents), then the behaviours that we want to shape will be actions in the world, not just prompt completions. We should not expect learning from human feedback, such as RLHF, to work well in such cases and the costs of inadequate alignment are likely to be greater. So, while learning from human feedback, such as RLHF or even Reinforcement Learning with AI Feedback (RLAIF), constitutes significant research achievements that are worth building on, we should also pursue other approaches to value implementation, through collaborative investigation drawing on different fields. These may include data curation and model unlearning, and implementing values in pre-training. We encourage exploration of how language models' competence with moral concepts can be operationalised to support more generalisable moral reasoning. High-level reasoning and planning capabilities are important constituents of responsible moral agency. Only agents that can plan can be consistent. Only agents with high-level reasoning capabilities can make complex value tradeoffs. So, while enhanced model capabilities will increase risk, they might also increase resources for successful value implementation.

Some obvious obstacles threaten progress in value implementation. Central challenges in model alignment include: Reward hacking, reward tampering, and specification gaming; the problem of supervising systems that are substantially more capable than humans; deceptive alignment, i.e.,

the possibility that models might appear to conform to intended values in training but depart from that in use.

Interdisciplinary collaboration faces obvious cold start problems widely discussed elsewhere[1]. In AI research, more resources are spent on advancing capabilities than on alignment, and funding for technical alignment dwarfs funding for sociotechnical work. AI companies have few social scientists and collaborate too infrequently with academia.

Progress is clearly possible, and will be marked by the following:

- Compelling answers to novel normative questions raised by advanced AI systems.
- Public and political education on the impacts of AI systems on more familiar values.
- Substantive political debate at international and domestic level over the future of AI.
- Innovation in participatory design by AI labs.
- More performant approaches to value implementation drawing on multiple fields.
- Better evaluations incorporating multidisciplinary approaches.
- Robust criteria for determining when value implementation has failed and AI systems are too unsafe to release.

1 https://nap.nationalacademies.org/catalog/26507/fostering-responsible-computing-research-foundations-and-practices

SCAI QUESTION 7

SCAI QUESTION 7

EQUITABLE ACCESS, CONTROL & FAIR COMPETITION

Where in the AI ecosystem should we ensure equitable access, control and fair competition? How should we address these concerns?

Context & Assumptions

Recent developments in AI have demonstrated tremendous potential to have both positive and negative impacts on society, organisations and individuals. AI systems currently rely upon large compute, advanced models, and extensive training data. These capabilities are inequitably distributed and result in a concentration of power. This in turn confers agency on specific actors, who may have goals that are misaligned with broader societal objectives. Areas of misalignment include an adequate recognition of risks, creating systems which are safe, and using AI to achieve social benefit or public good, rather than in support of profit maximisation. On the last, we note that the boundary between commercialization and basic research is not distinct. One of the assumptions driving research work, which may not be well founded, is that research can feed into a product that will in turn generate revenue and other resources to feed back into research. However, the majority of this research takes place in proprietary labs in companies that ship products and have profit as their central motivation.

Question

Where in the AI ecosystem should we ensure equitable access, control and fair competition? How should we address these concerns?

Concerns with concentration of power apply to a range of issues, including but not limited to price, quality, volatility, restrictions in access (with particular attention to which communities might be marginalised), restrictions in developing capability (e.g., training or hardware limitations) and ability to shape outcomes/output. There are also emerging externalities (e.g., situations of great individual benefit which result in collective harm) which may be exacerbated by such concentrations. We also realise there are unrecognised benefits to more open access to both the resources required to build AI models as well as to AI models themselves. These include broader economic growth across sectors, and broader perspectives in building and deployment. At the same time, there are some things that should remain closed or not be broadly accessible, e.g., PII and sensitive information, healthcare data.

Indicators of Progress

Given these considerations, we recommend two pathways to mitigate the effects of such concentration of power. First, develop a more democratic system that enables broader access to the key resources necessary to develop these technologies (i.e., lower barriers to entry for new entrants). This allows a wider range of actors of varying size and capability to the field, preventing or slowing concentration of power. Second, develop regulatory and non-regulatory strategies for the reduction of the harm in situations of power concentration. Non-regulatory strategies could include encouraging norms that support desired behaviours, such as transparency in

model-building to manage risk. These strategies are not unfamiliar. They have been used in the past to regulate other industries with similar potential impact, such as public utilities and telecommunications. For both pathways, potential areas for intervention, or chokepoints, include compute asymmetries (e.g., compute fabs, chip architecture, and optimisation for specific labs/models by chipmakers), and datasets (e.g., lack of representative datasets in underrepresented languages).

Some challenges to operationalising these strategies might be deeply held ideological beliefs about how the market should be structured, and risk tensions and tradeoffs (e.g., limited vs broader perspectives, more vs less control). Evolving use and emerging threats also require nimbleness in adapting regulation.

Indicators of success would include:

- Emergence of a mix of independent AI providers at different scales throughout the market i.e. both small and large firms.
- Large firms are well-regulated to minimise negative impact.
- Regulation differentiates between applications based on implications of their use (e.g., nuclear power vs nuclear weapons).
- Policy makers consider a clear framework when considering how to regulate different sectors/areas of the AI stack.

SCAI QUESTION 8

SCAI QUESTION 8

TRANSFORMING EDUCATION

How can we use AI to enhance the effectiveness, efficiency, and accessibility of education across societies around the world?

Context & Assumptions

Improvement of human capital through education is foundational across society and economy. There is great potential for AI to enhance the efficiency, effectiveness and accessibility across the entire educational ecosystem. This spans students, parents, teacher training, content creation, social interaction, curriculum design and delivery, evaluation and certification of results.

An AI-enhanced education ecosystem can cater to a range of needs across countries and contexts. In resource-constrained societies, this could mean the difference between education or no education at all. In other cases, knowledge and skills education could be enhanced to free up resources to address the more "human" aspects of education, such as critical thinking, creativity, empathy, social development etc, as well as focusing on student-teacher interactions. New models of human-AI collaboration could be explored, such as in using AI tools to assist teachers in characterising and evaluating the students and their progress, so as to challenge them appropriately and foster learning and personal growth. Across industry, appropriate interventions

could improve upskilling and the adoption of new tools and capabilities, enhancing productivity and introducing new value.

This approach relies on several assumptions. At the core, good access to infrastructure and connectivity are key; every student needs access to a mobile phone and a TV, possibly a keyboard. On the AI side, we assume the current state-of-the-art for AI, i.e. there are few unsolved scientific AI challenges that prevent us from making progress in this area and we can see payoff and test product-market-fit from the start.

Question

How can we design and implement an AI-enhanced, open education ecosystem, with a sustainable mechanism for participation, with the ambition of making education more efficient, effective, and accessible for local communities and global society?

Doing so will also enable us to maximise human capital across communities around the world, starting with maximising every student's potential.

Indicators of Progress

Given the complexity of education and the contextual needs of communities and corporations, a systems approach is necessary. This recognises that opportunities and incentives differ across life phases–elementary school, tertiary education, workforce reskilling, and more. Moreover, we need to address all elements of the value chain: students, teachers, curriculum optimization and

delivery, content creation, evaluation, certification, accessibility and social interactions. The following are non-exhaustive illustrations:

- **Students** can benefit from having an always-available tutor, and benefit from immediate feedback and personalised, adaptive content.

- **Parents** care about supporting and monitoring their childrens' progress. AI can help them support them, such as in pointing out areas of improvement and suggest actions.

- **Teachers** will benefit from spending less time on routine tasks such as grading, and can devote more time to foster creativity among students.

- **Content capture and creation** is time consuming, thus a major hurdle for educators to share courses. AI can be used to automatically transcribe and generate material.

- **Curriculum design and optimization** currently relies on the experience of lecturers, drawn from hundreds of students. A data-driven approach based on 100x more student expereiences can tailor curriculum for better learning, customised to individual students.

- **Evaluation** involves certification (of achieved degree), challenge (for progress) and control (of education objectives). Game mechanisms combined with empirical data can motivate students through suitably challenging problems and tests.

- **Accessibility** tools are typically costly (subtitles, speed, visual aids); these can be added through AI tools at minimal cost.

- **Student interaction** can be enhanced by social recommendations (student / tutor pairings, peer groups) and content moderation.

There is significant opportunity for productivity gains. These dividends can be used to increase the number of students trained, and/or the quality of their education, for a more cost-effective education system.

To achieve this, we must overcome several challenges, including:

- Overcoming accuracy issues such as hallucinations.
- Integration with existing infrastructure such as Google classroom.
- Interoperable, universal software interfaces that allow for whole system optimisation (future-proofing vs. ease of use vs. adoption), and that allows for meaningful auditability by humans; text-based interfaces are likely most suitable.
- Educator capabilities in using AI in teaching.
- Contextually-aligned curriculum and delivery depending on cultural norms and needs.
- Having strong enough economic incentives to facilitate diverse offerings and interoperability in an AI-enhanced education ecosystem.
- Regulatory and commercial acceptance of certificates obtained by AI-enhanced education (in particular for self-study and non-traditional pathways).
- Social acceptance of the notion of AI-enhanced education itself, by stakeholders such as educators, parents, students, companies, regulators.

SCAI QUESTION 9

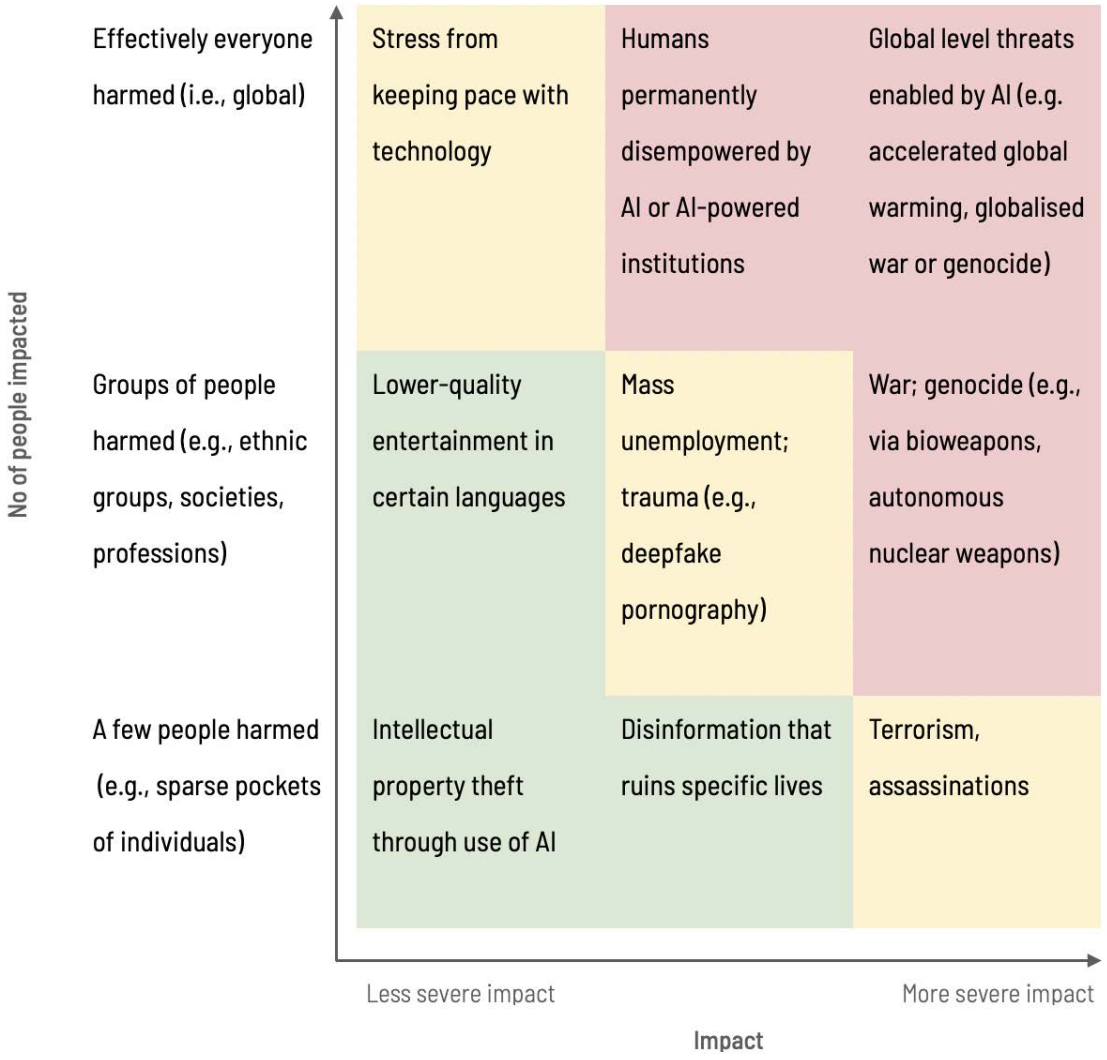
SCAI QUESTION 9

MITIGATING CATASTROPHIC RISKS & ONGOING HARMS

How can we mitigate the catastrophic risks and ongoing harms arising from AI, recognising that there are diverse opinions on the severity, probability, time sensitivity, and recoverability of these risks and harms?

Context & Assumptions

We recognise there is a range of views on what are the risks and harms that can arise, and their severity, probability, time sensitivity, and recoverability. For instance, here are some potential risks and our estimates on their time scales:



Types of potential catastrophic risks and harms	Time-scale of effect
Widespread social harm (e.g., loss of trust or trustworthiness in institutions, electoral dysfunction; employment challenges)	Already happening
AI-assisted cyber-risk	Already happening at some scale
Lethal autonomous weapons disasters	Already happening at some scale
AI-assisted bioweapons and accidents	Increasingly feasible now, and in need of greater attention
AI-assisted nuclear command and control malfunction	Some media reports suggest relevant discussions between some countries
AI-driven economic collapse (e.g., the 2010 Flash Crash at a much larger scale; mass unemployment)	Emerging / plausible within the next few years
AI-driven environmental destruction (e.g., exponentially accelerated pollution or resource consumption)	Plausible to begin within a decade or two

Risks and harms from AI can arise from various sources. They may occur by accident, intentionally, or due to willful indifference by the different stakeholders. They may also occur at the systemic level where no party is responsible or accountable when such risks and harms happen (e.g., widespread irreversible addiction to a technology that no one entity in particular is responsible for developing).

An assumption behind this question is that there may be warning signs, and it would be valuable to actively look for them. Catastrophic harms can also result even without Artificial General Intelligence (AGI) since there are AI capabilities in narrow domains that could already lead to societal-scale catastrophic risks.

Question

How can we mitigate the catastrophic risks and ongoing harms arising from AI, recognising that there are diverse opinions on the severity, probability, time sensitivity and recoverability of these risks and harms?

If we are to understand this, we will also need a way to discuss and identify which risks and harms are considered catastrophic and deserving of more attention. For each such risk, we will need to answer the following questions: what are its warning signs (if possible)? Who should be entrusted to monitor for those signs? On what time scale might it happen? Who decides if the risk is worth taking, and how? And if we fail to avoid it completely, how can we mitigate its effects and recover from it, and at what cost?

Indicators of Progress

To avoid catastrophic harms from increasingly advanced AI, we should establish clear warning signs and thresholds in advance across multiple areas. These include indicators and thresholds pertaining to computing power, demonstrations of dangerous AI abilities, job loss, lawsuits,

expert testimonies from diverse disciplines, and proliferation of fake content, impersonations, and cyberattacks.

Comprehensive safety evaluation infrastructure and standards should be developed for stress testing mission critical systems before full deployment. Benchmarks and standards should be developed for a range of technical and social considerations. Best practices such as red-teaming should also be established and scaled up.

Robust oversight mechanisms are also needed, involving consultation with diverse experts as well as representatives of potential victim groups. Audits should be independent without conflicts of interest. And given the global impact possible, international oversight mechanisms may be warranted for the most powerful AI systems.

By defining indicators and responses in a systematic way ahead of time, we can monitor progress and risk to make proactive governance decisions before harms arise. The goal is to avoid the "boiling frog" by reacting only when problems become dire and harder to address.

Potential challenges arise in achieving consensus on what constitutes a “catastrophe”, assessing the likelihood of various catastrophic risks, and determining how far off they are on the horizon. As such, this complexity necessitates a range of approaches, demanding more people and resources than would be required to mitigate a single type of catastrophic risk.

SCAI QUESTION 10

SCAI QUESTION 10

COMBATING MIS/DISINFORMATION CAMPAIGNS

What are the appropriate speed bumps and incentives for content channels to reduce the negative impact of mis/disinformation campaigns?

Context & Assumptions

Mis/disinformation is an existing problem, but with the development of AI, we will see an increase in volume and sophistication. This can be socially corrosive and degrade shared trust between citizens and institutions. The pervasiveness and velocity of social media content distributed through content channels have created the conditions where a generation relies on these channels to shape their understanding of the world.

This problem is hard to address because the techniques for mis/disinformation (especially multi-step emotional manipulation) are hard to detect. While AI tools lower the cost and increase access for those seeking to generate and proliferate disinformation, we are nearing a point where we lack the ability to discern if the source of information is human or bot and distinguish between true and fake content.

Question

What are the appropriate speed bumps and incentives for content channels to reduce the negative impact of mis/disinformation campaigns?

- What are the technical solutions to support these speed bumps/incentives?
- What are the trade offs between public safety and freedom for content generation (including the freedom to misrepresent authority of fact)?
- How do trusted institutions maintain trust with citizens in a world of increased mis/disinformation?

Indicators of Progress

While there is no known method to fully solve this problem, mitigation measures are possible. We can consider:

- Establish a digital identity system to allow for tracking sources of information.
- Promote third-party services that monitor content channels to flag disinformation, and correct it through decentralised systems that can tackle this problem at scale.

- Establish legal requirements for content channels to label AI generated videos, as a temporary measure while enabling the widespread adoption of digital signatures embedded in hardware manufacturing, such as digital signatures in cameras to authenticate images.

- Increase public education and awareness in unknowingly spreading mis/disinformation.

An additional challenge faced by non-English speaking countries is the technical difficulty of detecting mis/disinformation in non-English languages, since many existing models are trained on and optimised for English datasets. Therefore, algorithms need to be trained on non-English data sets in order to accurately detect on global platforms. Another challenge is the cost of fact checking posts at scale, which tends to be much higher than AI generation of fake posts. This is exacerbated by AI generated content being disseminated at an increasingly low cost.

There are few known systems to monitor the flow of misinformation. We anticipate both the public and private sector will need to invest in R&D to fill the void. Law enforcement will need to expand their investigative toolkits. Given the nascence of the problem, global sharing of experiences would improve the learning curve. Potentially, there will be AI trained specifically to address mis/disinformation.

SCAI QUESTION 11

SCAI QUESTION 11

A FRAMEWORK FOR EFFECTIVE AI ADOPTION FOR SOCIAL GOOD

How can AI adopters effectively evaluate and apply AI models for social good?

Context & Assumptions

AI developers often focus on improving technology, while governments regulate AI to address societal risks like misinformation or crime. Yet, there is a gap in effectively integrating AI into social good applications by governments, Non-Governmental Organisations (NGOs), and social enterprises, and a lack of thorough evaluation to measure their real impact.

In the private sector, AI adoption is measured by revenue and costs, making it easier for adopters to assess impact. However, in the social sector, evaluating outcomes in areas like education, healthcare, or climate change is more complex and lacks sufficient financial and technical resources for analysis. An additional complication is when adopters pick up a successful use case from one sector or context and apply it to their own, without evaluating whether the model or outcomes are still relevant for their situation. This complexity means a higher reputational risk for AI companies and greater potential harm from poorly designed programs, placing more

responsibility on the AI industry for effective adoption in social sectors.

The risks of early AI adoption include wasted investments and negative outcomes for participants, alongside reputational damage and potential overregulation for the AI industry. With growing interest in AI for social good, it is vital for the industry and adopters to develop a framework focusing on learning, piloting, evaluating, and capacity building.

For example, AI can boost teacher productivity and student learning in education, improve patient outcomes in healthcare, and provide better farming recommendations for climate change. However, rapid implementation without proper impact assessment can lead to negative consequences like reduced learning outcomes, incorrect health advice, or crop losses.

Question

How can AI adopters effectively evaluate and apply AI models for social good?

How can we offer a sociotechnical framework to AI adopters that enables them to:

- Accurately assess various aspects of AI models for social good use cases, including the dependencies (such as access to computational resources), utility (like model readiness and alignment with the proposed use case), and appropriateness (for instance, determining if general-use models are suitable for the intended purpose)?

- Pilot and rigorously evaluate AI use cases to comprehend their true impact compared to existing programs, ensuring that AI is being effectively adopted for social good?

Important considerations:

- It is crucial to emphasise socio-economic, cultural, and other differences that might lead to unintended consequences or worsen inequalities and biases when applying general-purpose algorithms in social sectors like education or public healthcare.

- AI adoption should not be rushed. It is essential to first pilot and rigorously evaluate AI-based interventions in real-world settings.

- The AI industry needs to allocate financial resources to support the "Framework for Effective AI Adoption for Social Good" proposed here. This includes funding for accessible in-person and online training for social sector organisations on integrating AI into their programs effectively; for conducting thorough evaluations of use cases; and for sharing case studies that highlight both successes and failures in these applications.

Indicators of Progress

We suggest the following potential approach:

- A Framework for Effective AI Adoption for Social Good requires that adopters:

    - Collaborate with the AI developer to gain a clear understanding of the potentials and limitations of the AI models, including the data they were trained on, the values underlying that model, and relevance to the local context.

    - Engage with researchers who possess global insights into both effective social

program design and AI integration. Involve them from the design stage to develop a pilot for the AI-enhanced program.

- Initiate a concurrent, independent, and rigorous evaluation, such as a randomised control trial (RCT), of the AI pilot. This is to accurately measure its impact compared to existing methods, offering insights into what works, what does not, and why.

- Proceed to scale up the program only after integrating learnings from both the pilot and the concurrent evaluation. If the pilot is found ineffective, consider scaling down.

- Disseminate the insights gained from this pilot and its evaluation to others who could benefit from integrating AI into their programs. Share through blogs or other open-source platforms.

- Address the needs of governments, NGOs, development organisations, and social enterprises in the context of social good applications (not commercial organisations and applications). These entities often lack the technical or financial resources for the above analysis, which is more challenging as outcomes are not solely measured in terms of revenue and costs.

- Known challenges or obstacles to answering this question include:

  - Program implementers often lack access to the global knowledge that can offer valuable insights for effectively incorporating AI into social programs. Researchers, with better access to this information, can be crucial partners in this process.

- There is sometimes an underestimation of the importance of tailoring the program to the local context. It is also vital to refine any elements that did not perform as expected.

- Even when social sector adopters recognize the importance of these steps, they frequently lack the technical capacity or financial resources to design and execute such a pilot and its thorough evaluation effectively.

- Tech and AI companies mostly do not allocate resources to assist adopters of their technology for social good. It is crucial for these companies to support social sector adopters in navigating this framework, ensuring that only relevant technology is adopted, and that it is done so correctly and appropriately.

- Broad criteria for recognising progress in answering the question:

  - The AI industry and adopters broadly use and adapt this "Framework for Effective AI Adoption for Social Good".

  - The AI industry commits both financial and technical resources for:

    - Implementing a sufficient number of use cases across various sectors such as education, public healthcare, climate change, social protection, labour markets, and agriculture, and in diverse socio-economic and geographic contexts.

- ➢ Publishing open-source studies that summarise the experiences of early adopters, including evidence from rigorous evaluations of the above use cases. These studies should focus on understanding what works, what does not, and why.

- Building capacity of key actors in the social sector to effectively adopt AI in their programs. This can be via a combination of in-person training for government AI/IT departments and multilateral development banks (like the World Bank and Asian Development Bank), as well as virtual training for relevant NGOs worldwide.

- Over time, as tech and AI companies, adopters, and researchers gain a better understanding of when and why AI is an effective tool in social sector programs, we expect to see fewer products exited and fewer pilots considered unsuccessful.

SCAI QUESTION 12

SCAI QUESTION 12

METHODOLOGIES FOR AI SAFETY EVALUATION

How can we establish and uphold methodologies for AI safety evaluation?

Context & Assumptions

Although high level concerns and principles are largely agreed upon for the ethical and safe use of AI (see for example the UNESCO 2022 recommendation), societies have yet to develop standardised techniques for mitigating harm and auditing procedures for safety testing. Although structures for governance and regulation are the topic of another question, here we focus on ways to operationalise auditing and transparency procedures. We define safety as adhering to the expected functionality of a system and avoiding unacceptable outcomes which may be considered harmful to individuals. Harms include social and psychological harms, and harms to security, economic or democratic resilience. Concerns extend throughout the lifecycle of a product, including after it ceases to be offered. We consider transparency as a core component of safe systems, enabling the tracing of accountability through the design and use of AI systems. Algorithm transparency is not just a tool for safety, but also an outcome of safety auditing processes.

Potential pitfalls include differences between the data on which a model is trained and the lived experience of the humans (or ecosystem) in which it is deployed; inadequate understanding of users' needs and context; and unanticipated or creeping harms such as gradual increase in loneliness or loss of human autonomy.

Current purely technical performance evaluation of AI systems – including large-scale generative models – is inadequate to measure and attribute correctly any increase of harms over the pre-AI-system status quo. Today, major AI developers and deployers use simple metrics such as diversities of datasets and outcomes, but this approach to auditing does not allow for comprehensive evaluation of socio-economic harms. To do this, we firstly need a scientific, data-driven method to understand the baseline and intended outcome pre-deployment, and subsequently whether there is any increase in harm post-deployment. Secondly, critical systems engineering requires testing the quality and reliability of system outputs, which should consider the context in which the objectives of the system and expectations of users are defined. These should include users' response to and understanding of the systems, which is in turn dependent on sociotechnical considerations, including user education.

Question

How can we establish and uphold methodologies for AI safety evaluation?

Relevant corollary questions include: What process can we use to capture and ensure the measurement of suspected harms (measuring macro/society and micro/individuals and families)? What statistics do we need about deployment, uptake, and modifications to deployed systems in order to establish causality between design choices and potentially negative social

outcomes? For any AI system, who are the users that are affected by either individual model components, or by human actors together with the model? How can we design a transparency report that clearly states the expected users, intended outcomes, and anticipated candidate harms?

How do we divide responsibility between developers of AI component systems and deployers whose products interact directly with the end users? We assume here that transparency requirements on deployers should be passed through their supply chain – that is, deployers are responsible for sourcing adequately-tested AI components, and for having a clear reporting path back to developers if issues are discovered among the deployers' users; developers are then responsible for system redesign according to this feedback. Who establishes adequacy benchmarks per deployment sector, and how are these communicated back to developers and/or used by deployers to assess systems' readiness for deployment?

Can controlled auditing and simulation effectively estimate societal risks? If so, what are the standards and obligations for testing model predictions before system deployment? In what contexts should we design adversarial testing, and with what frequency should such checks be executed?

There are further questions regarding post-deployment safety. What categories of AI systems benefit from paid incentivising for the reporting of safety and security problems (e.g. bug bounties, ethical hacking for AI)? How do we create and enforce processes for ordinary users to report suspected problems and receive responses (e.g., explanations)? How do we aggregate such reports and attribute them to candidate causes (e.g., flaws or active compromise of specific foundation models)?

Answering these questions should contribute to deployment of safe AI; improve AI development and innovation; ensure companies, governments, and potentially civil society have access to

adequate information about each other to do their jobs; and set precedents for transparent governance, bottom-up "policing" and understanding of rights.

Indicators of Progress

We will witness progress through the following measures:

- Corporate and civic confidence in deploying AI.
- Transparency to the public (e.g. through clear documentation) on how context-dependent acceptable and unacceptable outcomes of AI systems are defined.
- New standards, reusable tests and/or procedures for constructing tests, which focus on measuring reliability and impact. The costs/ benefits of AI use should be broken down by sub-populations.
- Standard checklists and regular reports, to be reported in media and made available through national standards organisations. The reporting of AI harms should be reliable (e.g., non-spurious).

A possible approach, in analogy with cybersecurity, is institutions like an AI CERT (Computer Emergency Response Team) that will gather and analyse reported AI vulnerabilities and failures and work with model developers and deployers to continuously improve the safety of the AI ecosystem. Such institutions could help address the specific post-deployment questions above.

Challenges include:

- Multiple stakeholders with competing interests.
- Gaining reliable access (at least for trusted parties) to proprietary systems or confidential data, which are needed for auditing.
- Ensuring the veracity of documents achieved through such access and that test performance corresponds to every-day, real-time performance.
- Designing metrics and actionable methods to accurately quantify risks to safety. This will require us to establish baseline social characteristics (e.g. using the World Values Survey).

ACKNOWLEDGEMENTS

ACKNOWLEDGEMENTS

The community-based process for producing these questions is outlined in the foreword. For more information about the SCAI process and community, please see https://www.scai.gov.sg/.

Published 6 December 2023