1st International Workshop on Evaluating Progress in Artificial Intelligence - EPAI 2020
In conjunction with the 24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Canaries in Technology Mines: Warning Signs of Transformative Progress in AI

Carla Zoe Cremer[1] and **Jess Whittlestone**[2]

Abstract. In this paper we introduce a methodology for identifying early warning signs of transformative progress in AI, to aid anticipatory governance and research prioritisation. We propose using expert elicitation methods to identify milestones in AI progress, followed by collaborative causal mapping to identify key milestones which underpin several others. We call these key milestones 'canaries' based on the colloquial phrase 'canary in a coal mine' to describe advance warning of an extreme event: in this case, advance warning of transformative AI. After describing and motivating our proposed methodology, we present results from an initial implementation to identify canaries for progress towards *high-level machine intelligence* (HLMI). We conclude by discussing the limitations of this method, possible future improvements, and how we hope it can be used to improve monitoring of future risks from AI progress.

1 INTRODUCTION

Progress in artificial intelligence (AI) research has accelerated in recent years, and applications are already beginning to impact society [9][43]. Some researchers warn that continued progress could lead to much more advanced AI systems, with potential to precipitate transformative societal changes [13][19][21][27][39]. For example, advanced machine learning systems could be used to optimise management of safety-critical infrastructure [33]; advanced language models could be used to corrupt our online information ecosystem [31]; and AI systems could even gradually begin to replace a large portion of economically useful work [17]. We use the term "transformative AI" to describe a range of possible advances in AI with potential to impact society in large and hard-to-reverse ways [22].

Preparing for the future impacts of AI is challenging given considerable uncertainty about how AI systems will develop. There is substantial expert disagreement around when different advances in AI capabilities should be expected [11][19][34]. Policy and regulation will likely struggle to keep up with the fast pace of technological progress [12][15][42], resulting in either stale, outdated regulation or policy paralysis [3]. It would therefore be valuable to be able to identify 'early warning signs' of transformative AI progress, to enable more anticipatory governance, as well as better prioritisation of research and allocation of resources.

We call these early warning signs 'canaries', based on the colloquial use of the phrase 'canary in a coal mine' to indicate advance warning of an extreme event. Our use of this term takes its inspiration from an article by Etzioni [16], in which he stresses the importance of identifying such canaries. We want to take this suggestion seriously and propose a systematic methodology for identifying such canaries.

Many types of indicators could be of interest and classed as canaries, including: research progress towards key cognitive faculties (e.g., natural language understanding); overcoming known technical challenges (such as improving the data efficiency of deep learning algorithms); or improved applicability of AI to economically-relevant tasks (e.g. text summarization). What distinguishes canaries from general markers of AI progress (like those discussed in [32] or [36]) is that they indicate that particularly transformative impacts of AI may be on the horizon. Given that our definition of "transformative AI" is currently very broad, canaries are therefore defined *relative to* a specific form of transformative AI or impact. For example, we might identify canaries for human-level artificial intelligence; canaries for automation of a specific sector of work; or canaries for specific types of societal risks from AI. From a governance perspective, we are particularly interested in canaries which indicate that *faster* or *discontinuous* progress may be on the horizon, since the impacts of rapid progress would be especially difficult to manage once they begin to manifest. These therefore particularly warrant advanced preparation.

In what follows, we describe and discuss a methodology for identifying canaries of progress in technological research. We focus on AI progress, but believe this method, once trialled and tested, could be applied to other areas of technological development. We motivate and describe the methodological approach, which combines expert elicitation and causal mapping, before presenting one implementation of this methodology to identify canaries for progress towards *high-level machine intelligence* (HLMI). After discussing potential canaries for HLMI specifically, we discuss how to make use of canaries in monitoring of AI progress and suggest how the limitations of this methodology might be addressed in future iterations of this work.

2 METHODOLOGY

2.1 Background and Motivation

This work builds on two main bodies of existing literature: research on *AI forecasting*, and an emerging body of work on *measuring AI progress*.

The AI forecasting literature generally uses expert elicitation to generate probabilistic estimates for when different types of

[1] Future of Humanity Institute, University of Oxford, UK, email: carla.cremer@philosophy.ox.ac.uk
[2] Centre for the Study of Existential Risk, University of Cambridge, UK, email: jlw84@cam.ac.uk

advanced AI will be achieved [5][19][21][34]. For example, Baum et al. [5] survey experts on when four specific milestones in AI will be achieved: passing the Turing Test, performing Nobel quality work, passing third grade, and becoming superhuman. Both Müller and Bostrom [34], and Grace et al. [19] focus their survey questions around predicting the arrival of *high-level machine intelligence* (HLMI), which the latter define as being achieved when "unaided machines can accomplish every task better and more cheaply than human workers".

However, these studies have several limitations [6] that should make us cautious about giving their results too much weight. The experts surveyed in these studies have no experience in quantitative forecasting and receive no training before being surveyed, which likely renders their predictions unreliable [10][41].

Issues of reliability aside, these forecasting studies are also limited in what they can tell us about future AI progress. They have little to say about impactful milestones on the path to advanced AI, let alone early-warning signs. Experts disagree substantially about when capabilities will be achieved [11][19] and without knowing who (if anyone) is more accurate in their predictions, these forecasts cannot easily inform decisions and prioritisation around AI today. Quantitative expert elicitations like these also do not tell us *why* different experts disagree, what kinds of progress might cause them to change their judgements, or what aspects they in fact do agree upon. While broad probability estimates for when advanced AI might be achieved are interesting, they tell us little about the path from where we are now, or what could be done today to shape the future development and impact of AI.

At the same time, several research projects have begun to track and measure progress in AI [7][23][36]. These projects focus on a range of indicators relevant to AI progress, but do not make any systematic attempt to identify which markers of progress are most *important* for anticipating potentially transformative AI. Given limited time and resources for tracking progress in AI, it is crucial that we find ways to prioritise those measures that are most relevant to ensuring societal benefits and mitigating risks of AI.

The approach we propose in this paper aims to address the limitations of both work on AI forecasting and on measuring progress in AI. In a sense, the limitations of these two bodies of work are complementary. The AI forecasting literature focuses on anticipating the most extreme impacts and advanced progress in AI, but is unable to say much about the warning signs or kinds of progress that will be important in the near future. AI measurement initiatives, conversely, monitor near-future progress in AI, but with little systematic prioritisation or reflection on what progress in different areas might mean for society and governance. What is needed, building on work in both these areas, are attempts to identify areas of progress today that may be particularly important to pay attention to, *given* concerns about the kinds of transformative AI systems that may be possible in future. Progress in these areas would serve as crucial warning signs - canaries, as well call them - suggesting more advance preparation for future AI systems and their impacts is needed.

We believe that identifying canaries for transformative AI is a tractable problem and therefore worth investing considerable research effort in today. In both engineering and cognitive development, capabilities are achieved sequentially, meaning that there are often key underlying capabilities which, if attained, unlock progress in many other areas. For example, musical protolanguage is thought to have enabled grammatical competence in the development of language in homo sapiens [8]. AI progress so far has also arguably seen such amplifiers: the use of multi-layered non-linear learning or stochastic gradient descent arguably laid the foundation for unexpectedly fast progress on image recognition, translation and speech recognition [29]. By mapping out the dependencies between different capabilities or milestones needed to reach some notion of transformative AI, therefore, we should be able to identify milestones which are particularly important for enabling many others - these are our canaries. This is the general idea behind our approach to identifying canaries, outlined in more detail in the following sections.

2.2 Proposed Methodology

The proposed methodology can be used to identify 'canaries' for any transformative event. In the case of AI, the focus might be on a transformative technology such as HLMI or AGI, a transformative application such as flexible robotics, or a transformative impact such as the automation of at least 50% of jobs.

Given a transformative event, our methodology has three main steps: (1) identifying key milestones towards the event; (2) identifying dependency relations between these milestones; and (3) identifying milestones which underpin many others as 'canaries'.

2.2.1 Identifying key milestones using expert elicitation

Like other studies of AI progress, we rely on expert elicitation throughout this process. However, the reliability of expert elicitation studies depends on the proper selection and use of expertise. Though there are inevitable limitations of using any form of subjective judgement, no matter how expert, these limitations can be minimised with careful selection of experts and questions.

We suggest carefully selecting experts with varied expertise relevant to the chosen question. For example, for identifying milestones towards human-level AI, the cohort should include experts in machine learning and computer science but also cognitive scientists, philosophers, developmental psychologists, evolutionary biologists, and animal cognition experts. This diverse group would bring together expertise on the current capabilities and limitations of AI, with expertise on key milestones in human cognitive development and the order in which cognitive faculties develop. We also encourage careful design and phrasing of questions to enable participants to make best use of their expertise. For example, rather than asking experts to identify specific milestones towards human-level AI, which is a question for which they are not trained, we might ask machine learning researchers about the limitations of the methods they use every day, or ask psychologists what important human capacities they see lacking in machines.

There are several different methods available for expert elicitation: including surveys, interviews, workshops and focus groups, each with advantages and disadvantages [2]. Interviews provide greater opportunity to tailor questions to the specific expert, but can be extremely time-intensive compared to surveys, making it difficult to consult a large number of experts. If possible, some combination of the two may be ideal: using carefully selected semi-structured interviews to elicit initial milestones, followed-up with surveys with a much broader group to validate which milestones are widely accepted as being key.

2.2.2 Mapping dependencies between milestones using causal graphs

The second step of our methodology involves convening experts to identify dependency relations between identified milestones: that is, which milestones may underpin, lead to, or depend on which others. Experts should be guided in generating *directed causal graphs* to represent perceived causal relations between milestones [35]. Causal graphs show causal links between elements of a system, represented as nodes (elements) and arrows (causal links). A directed positive arrow from A and B indicates that A has a positive causal influence on B. Such causal maps have been used to support decision-making, structure knowledge, and improve visualisation of complex scenarios [20][25][26][28] and are particularly useful for exploring and understanding possible futures, rather than aiming to predict a single future [26]. They are easily modified and constructed collaboratively, and therefore are well-suited to helping us structure expert knowledge on dependencies between different technological milestones.

Fuzzy cognitive maps (FCMs), a specific type of causal graph, may be a particularly useful method for our purposes. FCMs capture all benefits of causal mapping but can be extended into a quantitative model, and thus lend themselves to computerised simulations [25]. This will not always be necessary, but given that our proposed method is applicable to many contexts, a flexible model is desirable. FCMs are well able to document non-linear interactions and experts' mental models of causal interactions because they can handle imprecise causal links. The variables (nodes) can take any state between 0 and 1 (hence 'fuzzy'), indicating the extent to which the variable is 'present'. When a variable changes its state, it affects all concepts that are causally dependent on it. FCMs have been used successfully in environmental science [20][38], strategic planning [30], and other areas [25], and have been recommended for use in futures studies, forecasting, and technology road mapping [1][26].

In a workshop format, experts should be given brief training in causal graph methods or FCMs, and then break into groups to discuss dependencies between milestones. Each group should then collaboratively construct a directed causal graph or FCM to represent these relationships. Groups should be formed so as to maximise the variation of expertise in each group.

2.2.3 Identifying canaries from causal graphs.

Finally, the resulting causal graphs can be aggregated and analysed to identify canaries, by identifying the nodes with the highest number of outgoing arrows.

The aggregation process should first focus on identifying commonalities between all graphs which can be shared in the final graph. Substantive disagreements may remain, which can be the subject of mediated discussion, with a voting process to decide on final aspects of the graph.

Experts then identify nodes which they agree have significantly more outgoing nodes (some amount of discretion from the experts/conveners will be needed to determine what counts as 'significant'). Since nodes with a high density of outgoing arrows represent milestones which underpin many others, progress on these milestones can act as 'canaries', indicating that we may see further advances in many other areas in the near future. These canaries can therefore act as early warning signs for more rapid and potentially discontinuous progress, as well as for new applications becoming ready for deployment (depending on which exact capabilities they are likely to unlock).

3 IMPLEMENTATION: CANDIDATE CANARIES FOR HLMI

We describe a partial implementation of the proposed method to identify canaries for achieving high-level machine intelligence (HLMI). We define HLMI here as an AI system (or collection of AI systems) that performs at the level of an average human adult on key cognitive measures required for economically relevant tasks.[2] We interviewed experts about the limitations of current deep learning methods from the perspective of achieving HLMI, and used the findings to construct a causal graph of milestones. This allowed us to identify candidate canary capabilities. The results must be understood as preliminary, because the causal graphs were developed just by the authors, not a cohort of experts, and so have limited validity. However, this initial demonstration and preliminary findings will form the basis for a full study with a broader range of experts in future.

3.1 Expert elicitation to identify milestones

To identify key milestones for achieving HLMI, we interviewed 25 experts (using both a non-probabilistic, purposive sampling method and stratified sampling method, as described by [12] in chapter six). The sample covered experts in machine learning (9), computer science with specialisation in AI (5), cognitive psychology (2), animal cognition (1), philosophy of mind and AI (3), mathematics (2), neuroscience (1), neuro-informatics (1), engineering (1). Interviewees came from both academia and industry, and were deliberately selected to be at a variety of career stages.

We conducted individual, semi-structured interviews, with a set of core questions and themes to guide more open-ended discussion. Semi-structured interviews use an interview guide with core questions and themes to be explored in response to open-ended questions to allow interviewees to explain their position freely [24]. Initial questions included: what do you believe deep learning will never be able to do? Do you see limitations of deep learning that others seem not to notice? In response to these and similar questions tailored to the interviewee's specific expertise, they were asked to name what they thought were the biggest limitations of current deep learning methods, from the perspective of achieving HLMI.

All named limitations were collated, with shortened explanations, and translated into 'milestones', i.e. capabilities experts believe deep learning is yet to achieve on the path to HLMI. Table 1. shows all milestones based on limitations, named by interviewees. Because we have maintained each interviewee's preferred terminology, several of the milestones listed may turn out to refer to the same or highly similar problems.

Table 1. Limitations of deep learning as perceived and named by experts

[2] We use this definition, adapted from Grace et al., to highlight that what is key for saying HLMI has been reached is that AI has the *cognitive ability* to perform every task better than humans workers, not that it is in practice deployed to do so.

Causal reasoning: the ability to detect and generalise from causal relations in data.	**Common sense:** having a set of background beliefs or assumptions which are useful across domains and tasks.
Meta-learning: the ability to learn how to best learn in each domain.	**Architecture search**: the ability to automatically choose the best architecture of a neural network for a task.
Hierarchical decomposition: the ability to decompose tasks and objects into smaller and hierarchical sub-components.	**Cross-domain generalization**: the ability to apply learning from one task or domain to another.
Representation: the ability to learn abstract representations of the environment for efficient learning and generalisation.	**Variable binding:** the ability to attach symbols to learned representations, enabling generalisation and re-use.
Disentanglement: the ability to understand the components and composition of observations, and recombine and recognise them in different contexts.	**Analogical reasoning:** the ability to detect abstract similarity across domains, enabling learning and generalisation.
Concept formation: the ability to formulate, manipulate and comprehend abstract concepts.	**Object permanence:** the ability to represent objects as consistently existing even when out of sight.
Grammar: the ability to construct and decompose sentences according to correct grammatical rules.	**Reading comprehension:** the ability to detect narratives, semantic context, themes and relations between characters in long texts or stories.
Mathematical reasoning: the ability to develop, identify and search mathematical proofs and follow logical deduction in reasoning.	**Visual question answering:** the ability to answer open-ended questions about the content and interpretation of an image.
Uncertainty estimation: the ability to represent and consider different types of uncertainty.	**Positing unobservables:** the ability to account for unobservable phenomena, particularly in representing and navigating environments.
Reinterpretation: the ability to partially re-categorise, re-assign or reinterpret data in light of new information without retraining from scratch.	**Theorising and hypothesising:** the ability to propose theories and testable hypotheses, understand the difference between theory and reality, and the impact of data on theories.
Flexible memory: the ability to store, recognise and retrieve knowledge so that it can be used in new environments and tasks.	**Efficient learning**: the ability to learn efficiently from small amounts of data.
Interpretability: the ability for humans to interpret internal network dynamics so that researchers can manipulate network dynamics.	**Continual learning:** the ability to learn continuously as new data is acquired.
Active learning: the ability to learn and explore in self-directed ways.	**Learning from inaccessible data:** the ability to learn in domains where data is missing, difficult or expensive to acquire.
Learning from dynamic data: the ability to learn from a continually changing stream of data.	**Navigating brittle environments:** the ability to navigate irregular, and complex environments which lack clear reward signals and short feedback loops.
Generating valuation functions: the ability to generate new valuation functions immediately from scratch to follow newly-given rules.	**Scalability:** the ability to scale up learning to deal with new features without needing disproportionately more data, model parameters, and computational power.
Learning in simulation: the ability to learn all relevant experience from a simulated environment.	**Metric identification:** the ability to identify appropriate metrics of success for complex tasks, such that optimising for the measured quantity accomplishes the task in the way intended.
Conscious perception: the ability to experience the world from a first-person perspective.	**Context-sensitive decision making:** the ability to adapt decision-making strategies to the needs and constraints of a given time or context.

It is worth noting there are apparent similarities and relationships between many of these milestones. For example, representation: the ability to learn abstract representations of the environment, seems closely related to variable binding: the ability to formulate place-holder concepts. The ability to apply learning from one task to another, cross-domain generalisation, seems closely related to analogical reasoning. Further progress in research will tell which of these are clearly separate milestones or more closely related notions.

3.2 Causal graphs to identify dependencies between milestones

Having identified key milestones, we explore dependencies between them using fuzzy cognitive maps (FCM). We focus on how capabilities enable, not inhibit, other capabilities, which means we use only positive influence arrows. FCMs are particularly well-suited to representing the uncertainty inherent in this analysis, as it assumes that each arrow could have a weight to represent varying levels of strength. In this analysis we have not specified the weights on connections, but adding these weights could be trialled with experts in the future.

A previous survey [5] suggests that this endeavour is a highly uncertain one, finding that many different relationships between AI milestones seem plausible to experts. Our analysis does not claim nor aim to resolve this disagreement, but instead shows only one out of many possible mappings, to illustrate the use and value of FCMs in AI progress monitoring.

We use the software VenSim (vensim.com) to illustrate the hypothesised relationships between perceived milestones in Figure 1. For example, we hypothesise that the ability to formulate, comprehend and manipulate abstract concepts may be an important prerequisite for the ability to account for unobservable phenomena, which is in turn important for reasoning about causality. A positive influence arrow does not mean that achieving one milestone *necessarily* leads to another, but rather that progress on the first

milestones increases the likelihood of progress on other arrows it points to.

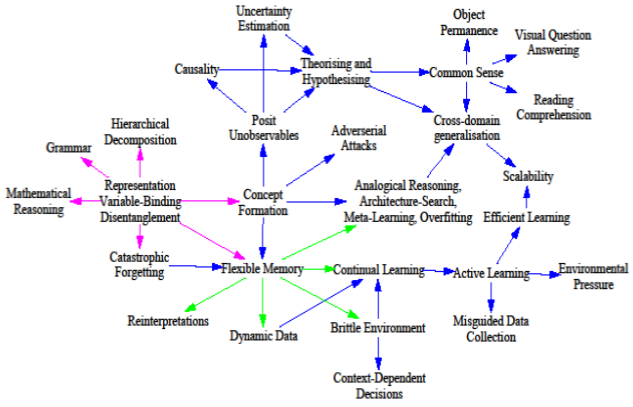

Figure 1. Cognitive map of dependencies between milestones collected in expert elicitations. Arrows coloured in pink and green indicate capabilities that have significantly more outgoing arrows.

This map was constructed by the authors, and is therefore far from definitive or the only possible way of representing dependencies between capabilities. However, this initial map does provide an important illustration of the kind of output this methodology should aim to achieve, and generates some initial hypotheses for relationships between milestones.

3.3 Candidate Canary Capabilities

Based on this causal map, we can identify two candidates for canary capabilities. The capabilities with the most outgoing arrows are:

Symbol-like representations: the ability to construct abstract, discrete and disentangled representations of inputs, to allow for efficiency and variable-binding. We hypothesise that this capability underpins several others, including grammar, mathematical reasoning, concept formation, and flexible memory.

Flexible memory: the ability to store, recognise, and re-use knowledge. We hypothesise that this ability would unlock many others, including the ability to learn from dynamic data, the ability to learn in a continual fashion, and the ability to learn how to learn.

We therefore tentatively suggest that these are two important capabilities to track progress on from the perspective of anticipating HLMI. We discuss one such capability, flexible memory, in more detail below.

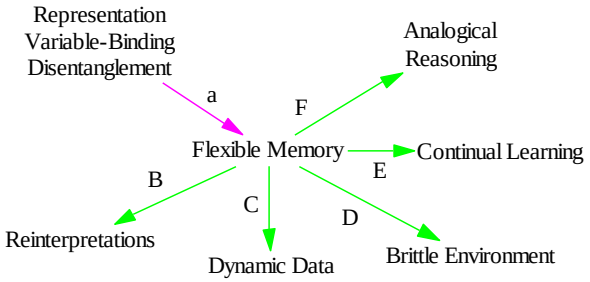

Figure 2. Extract of Figure 1, showing one candidate canary capability.

Flexible memory, as described by experts in our sample, is the ability to recognize and store reusable information, in a format that is flexible so that it can be retrieved and updated when new knowledge is gained. We explain the reasoning behind the labelled arrows in figure 2:

- (a): compact representations are a prerequisite for flexible memory since storing high-dimensional input in memory requires compressed, efficient and thus abstract representations.
- (B): the ability to reinterpret data in light of new information likely requires flexible memory, since it requires the ability to retrieve and alter previously stored information.
- (C) and (E): to make use of dynamic and changing data input, and to learn continuously over time, an agent must be able to store, correctly retrieve and modify previous data as new data comes in.
- (D): in order to plan and execute strategies in brittle environments with long delays between actions and rewards, an agent must be able to store memories of past actions and rewards, but easily retrieve this information and continually update its best guess about how to obtain rewards in the environment.
- (F): analogical reasoning involves comparing abstract representations, which requires forming, recognising, and retrieving representations of earlier observations.

Progress in flexible memory therefore seems likely to unlock or enable many other capabilities important for HLMI, especially those crucial for applying AI systems in real environments and more complex tasks. These initial hypotheses should be validated and explored in more depth by a wider range of experts.

4 DISCUSSION

4.1 Advantages

We believe the proposed method for identifying canaries has many strengths and could be applied to a broad range of important questions about transformative AI systems and impacts. The general methodology of using expert elicitation to identify milestones and then causal mapping to elucidate dependencies between those milestones is extremely flexible, meaning it could be applied beyond AI to other fields of science and technology progress. The method can also be adapted to the preferred level of detail for a given study: causal graphs can be made arbitrarily complex [18] and can be analysed both quantitatively and qualitatively. With this method, it is possible to combine different types of expertise relating to milestones: including well-understood technical limitations of current methods, with informed speculation about unknown capabilities that may be important prerequisites to some transformative event. With early warning signs we can track progress towards canary milestones, or directly prepare for the transformative events that follow after it.

4.2 Uses

We envision that this methodology could be used to identify warning signs for a number of important potentially transformative events in AI progress, such as foundational research break-throughs, the use of AI to automate scientific research, or the automation of tasks that affect a wide range of jobs.

Once canaries have been identified for some transformative event, there are numerous ways we might use them to improve preparation for its impact, including by:

- Automating the collection, tracking and flagging of new publications relevant to canary capabilities, and building a database of relevant publications (perhaps similar to that described by [40]);
- Generating metrics and benchmarks for evaluating progress on canary capabilities;
- Using prediction platforms such as Metaculus (ai.metaculus.com) to track and forecast progress on canary capabilities;
- Conducting more focused expert elicitation, for example periodically consult experts on their updated forecasts (in the form of cumulative probability estimates) for when different milestones are achieved, or when they are presented with updated progress metrics on canary capabilities;
- Conducting more in-depth research to empirically and theoretically investigate hypothesised relationships between milestones: for example, to what extent do improvements in memory structures lead to empirical improvements in performance in brittle environments?
- Conducting more in-depth research on the societal and governance implications of achieving canary milestones, and preparing governance responses for these milestones ahead of time.

4.3 Limitations and future directions

This methodology nonetheless has some limitations which further iterations could seek to improve on. There may be a fundamental trade-off between the benefits of consulting a large, diverse group of experts - enabling more thorough and robust identification of relevant milestones - and the feasibility of reaching agreement upon a single causal map, and therefore agreeing upon canaries. Relatedly, if uncertainty about milestones is too high, it may be difficult for experts to agree on a single causal map or candidates for canaries: finding questions where there is enough uncertainty for this process to be useful, but not so much uncertainty that no agreement can be reached, may be a challenge in some cases. It will also be important to recognise any potential limitations of the specific sample of experts involved in the process: recognising that machine learning researchers may be biased towards emphasising the importance of areas they themselves work on, for example, or that non-computer scientists may often lack a full understanding of what current systems can and cannot do.

In using FCMs to generate causal maps, it is not clear what level of detail and quantitative analysis will be most useful. In the implementation described here, we hypothesised relationships at a high level of abstraction and without quantitative analysis, due to the high level at which experts highlighted limitations in the first stage. The higher the level of abstraction, the more uncertain the mapping will be and the less useful it may be to indicate weights. It would be valuable for future work to explore various levels of abstractions, including a more detailed and quantitative analysis using more clearly-defined technical milestones, which could result in more precise forecasts and hypotheses.

Finally, it is important to note that attempts to anticipate and understand progress in AI (or any other technology) are not independent of that progress itself. Better understanding of key milestones towards AGI, HLMI, or some other notion of transformative AI, does not just improve our ability to anticipate that progress, but may also improve our ability to *make* progress towards transformative AI. We must therefore be cautious in identifying 'canary' capabilities, to consider the potential risks of making progress on these capabilities, and to communicate and encourage consideration of these risks to those researchers driving forward AI development.

REFERENCES

[1] M. Amer, A. Jetter, and T. Daim, 'Development of fuzzy cognitive map (FCM)-based scenarios for wind energy', *International Journal of Energy Sector Management*, vol. 5, no. 4, pp. 564–584, 2011. doi: 10.1108/17506221111186378.

[2] B.M. Ayyub, 'Elicitation of expert opinions for uncertainty and risks', *CRC press*, 2001.

[3] S. Ballard and R. Calo, 'Taking Futures Seriously: Forecasting as Method in Robotics Law and Policy', Draft available at: https://robots.law.miami.edu/2019/wp-content/uploads/2019/03/Calo_Taking-Futures-Seriously.pdf, 2019.

[4] S. Baum, B. Goertzel, and T.G. Goertzel, 'How long until human-level AI? Results from an expert assessment', *Technological Forecasting and Social Change*, 78(1), 185-195, 2011.

[5] S. Beard, T. Rowe, and J. Fox, 'An analysis and evaluation of methods currently used to quantify the likelihood of existential hazards', *Futures*, 115, 102469, 2020.

[6] N. Benaich and I. Hogarth, 'State of AI Report 2019', available at https://www.stateof.ai/, 2019.

[7] C. Buckner and K. Yang, 'Mating dances and the evolution of language: What's the next step?', *Biology & Philosophy*, vol. 32, 2017, doi: 10.1007/s10539-017-9605-z.

[8] K. Crawford, R. Dobbe, T. Dryer, G. Fried, B. Green, E. *et al*., 'AI Now 2019 Report', New York: AI Now Institute, 2019.

[9] W. Chang, E. Chen, B. Mellers and P. Tetlock, 'Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments', *Judgement and Decision Making* , vol. 11, no. 5, pp. 509-526, 2016.

[10] C. Z. Cremer, 'Deep Limitations? Examining Expert Disagreement Over Deep Learning', Manuscript in preparation, 2020.

[11] D. Collingridge, 'The Social Control of Technology', *St. Martin's Press* 1980.

[12] A. Dafoe, 'AI Governance: A Research Agenda', Governance of AI Program, The Future of Humanity Institute, The University of Oxford, Oxford, UK, 2018.

[13] M. DeJonckheere and L. Vaughn, 'Semistructured interviewing in primary care research: a balance of relationship and rigour', *Family Medicine and Community Health*, pp. doi: 10.1136/fmch-2018-000057, 2019.

[14] O. Etzioni, 'How to know if artificial intelligence is about to destroy civilization', *MIT Technology Review*, 2020.

[15] C. B. Frey and M. A. Osborne, 'The future of employment: How susceptible are jobs to computerisation?' *Technological Forecasting and Social Change,* vol. 114, pp. 254-280, 2017.

[16] S. Friel *et al*., 'Using systems science to understand the determinants of inequities in healthy eating', *PLoS ONE*, vol. 12, no. 11, p. e0188872, 2017. doi: 10.1371/journal.pone.0188872.

[17] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, and O. Evans, 'When will AI exceed human performance? Evidence from AI experts', *Journal of Artificial Intelligence Research*, 62, 729-754, 2018.

[18] S. R. J. Gray *et al*., 'Are coastal managers detecting the problem? Assessing stakeholder perception of climate vulnerability using Fuzzy Cognitive Mapping', *Ocean & Coastal Management*, vol. 94, pp. 74–89, 2014. doi: 10.1016/j.ocecoaman.2013.11.008.

[19] R. Gruetzemacher, 'A Holistic Framework for Forecasting Transformative AI', *Big Data and Cognitive Computing*, 3(3): 35, 2019.

[20] R. Gruetzemacher and J. Whittlestone, 'Defining and Unpacking Transformative AI', *arXiv preprint arXiv:1912.00747*. 2019.

[21] A. Haynes and L. Gbedemah, 'The Global AI Index Methodology', *Tortoise Media*, available at: https://www.tortoisemedia.com/intelligence/ai/, 2019.

[22] S. Jamshed, 'Qualitative research method-interviewing and observation', *Journal of Basic and Clinical Pharmacy*, vol. 5, no. 4, p. 87–88, 2014.

[23] A. Jetter, 'Fuzzy Cognitive Maps for Engineering and Technology Management: What Works in Practice?', in *2006 Technology Management for the Global Future - PICMET 2006 Conference*, Istanbul, Turkey, pp. 498–512, 2006. doi: 10.1109/PICMET.2006.296648.

[24] J. Jetter and K. Kok, 'Fuzzy Cognitive Maps for futures studies—A methodological assessment of concepts and methods', *Futures*, vol. 61, pp. 45–57, 2014, doi: 10.1016/j.futures.2014.05.002.

[25] H. Karnofsky, 'Some Background on our Views Regarding Advanced Artificial Intelligence', *available at*: *https://www.openphilanthropy.org/blog/some-background-our-views-regarding-advanced-artificial-intelligence*, 2016.

[26] B. Kosko, 'Fuzzy Cognitive Maps', *Int. J. Man-Machine Studies*, vol. 24, pp. 65-75, 1986.

[27] Y. LeCun, Y. Bengio and G. Hinton, 'Deep Learning', *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.

[28] K. C. Lee, H. Lee, N. Lee, and J. Lim, 'An agent-based fuzzy cognitive map approach to the strategic marketing planning for industrial firms', *Industrial Marketing Management*, vol. 42, no. 4, pp. 552–563, 2013, doi: 10.1016/j.indmarman.2013.03.007.

[29] H. Lin, 'The existential threat from cyber-enabled information warfare', *Bulletin of the Atomic Scientists*, 75(4), 187-196, 2019.

[30] F. Martínez-Plumed, S. Avin, M. Brundage, A. Dafoe, S.S. ÓhÉigeartaigh, and J. Hernández-Orallo, 'Accounting for the neglected dimensions of AI progress', *arXiv preprint arXiv:1806.00610*, 2018.

[31] M. Mohammadi, and A. Al-Fuqaha, 'Enabling cognitive smart cities using big data and machine learning: Approaches and challenges', *IEEE Communications Magazine*, 56(2), 94-101, 2018.

[32] V. C. Müller, and N. Bostrom, 'Future progress in artificial intelligence: A survey of expert opinion', in *Fundamental issues of artificial intelligence* (pp. 555-572). Springer, Cham, 2016.

[33] B. Newell and K. Proust, 'Introduction to Collaborative Conceptual Modelling', p. 20, 2012.

[34] R. Perrault, Y. Shoham, E. Brynjolfsson, J. Clark, J. Etchemendy *et al*., 'The AI Index 2019 Annual Report', AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA, 2019.

[35] K. Proust *et al*., 'Human Health and Climate Change: Leverage Points for Adaptation in Urban Environments', *IJERPH*, vol. 9, no. 6, pp. 2134–2158, 2012, doi: 10.3390/ijerph9062134.

[36] D. Reckien, 'Weather extremes and street life in India—Implications of Fuzzy Cognitive Mapping as a new tool for semi-quantitative impact assessment and ranking of adaptation measures', *Global Environmental Change*, vol. 26, pp. 1–13, 2014, doi: 10.1016/j.gloenvcha.2014.03.005.

[37] G. Shackelford, L. Kemp, C. Rhodes, L. Sundaram, S.S. ÓhÉigeartaigh *et al*., 'Accumulating evidence using crowdsourcing and machine learning: A living bibliography about existential risk and global catastrophic risk', *Futures*, 116, 2020. https://doi.org/10.1016/j.futures.2019.102508

[38] P. Tetlock and D. Gardner, 'Superforecasting, The Art and Science of Prediction', *London: Random House Books*, 2016.

[39] W. Wallach, 'A dangerous master: How to keep technology from slipping beyond our control', *Basic Books*, 2015.

[40] J. Whittlestone, R. Nyrup, A. Alexandrova, K. Dihal, and S. Cave, 'Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research', *London: Nuffield Foundation*, 2019.