Artificial Canaries: Early Warning Signs for Anticipatory and Democratic Governance of AI

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

[1] Future of Humanity Institute, University of Oxford (United Kingdom)
[2] Centre for the Study of Existential Risk, University of Cambridge (United Kingdom)

Received 17 January 2021 | Accepted 11 February 2021 | Published 24 February 2021

Abstract

We propose a method for identifying early warning signs of transformative progress in artificial intelligence (AI), and discuss how these can support the anticipatory and democratic governance of AI. We call these early warning signs 'canaries', based on the use of canaries to provide early warnings of unsafe air pollution in coal mines. Our method combines expert elicitation and collaborative causal graphs to identify key milestones and identify the relationships between them. We present two illustrations of how this method could be used: to identify early warnings of harmful impacts of language models; and of progress towards high-level machine intelligence. Identifying early warning signs of transformative applications can support more efficient monitoring and timely regulation of progress in AI: as AI advances, its impacts on society may be too great to be governed retrospectively. It is essential that those impacted by AI have a say in how it is governed. Early warnings can give the public time and focus to influence emerging technologies using democratic, participatory technology assessments. We discuss the challenges in identifying early warning signals and propose directions for future work.

Keywords

AI Governance, Forecasting, Anticipatory Governance, Participatory Technology Assessments.

DOI: 10.9781/ijimai.2021.02.011

I. Introduction

Progress in artificial intelligence (AI) research has accelerated in recent years. Applications are already changing society [1] and some researchers warn that continued progress could precipitate transformative impacts [2]–[5]. We use the term "transformative AI" to describe a range of possible advances with potential to impact society in significant and hard-to-reverse ways [6]. For example, future machine learning systems could be used to optimise management of safety-critical infrastructure [7]. Advanced language models could be used in ways that corrupt our online information ecosystem [8] and future advances in AI systems could trigger widespread labour automation [9].

There is an urgent need to develop anticipatory governance approaches to AI development and deployment. As AI advances, its impacts on society will become more profound, and some harms may be too great to rely on purely 'reactive' or retrospective governance.

Anticipating future impacts is a challenging task. Experts show substantial disagreement about when different advances in AI capabilities should be expected [10], [11]. Policy-makers face challenges in keeping pace with technological progress: it is difficult to foresee impacts before a technology is deployed, but after deployment it may already be too late to shape impacts, and some harm may already have been done [12]. Ideally, we would focus preventative, anticipatory efforts on applications which are close enough to deployment to be meaningfully influenced today, but whose impacts we are not already seeing. Finding 'early warning signs' of transformative AI applications can help us to do this.

Early warning signs can also help democratise AI development and governance. They can provide time and direction for much-needed public discourse about what we want and do not want from AI. It is not enough for anticipatory governance to look out for supposedly 'inevitable' future impacts. We are not mere bystanders in this AI revolution: the futures we occupy will be futures of our own making, driven by the actions of technology developers, policymakers, civil society and the public. In order to prevent foreseeable harms towards those people who bear the effects of AI deployments, we must find ways for AI developers to be held accountable to the society which they are embedded in. If we want AI to benefit society broadly, we must urgently find ways to give democratic control to those who will be impacted. Our aim with identifying early warning signs is to develop anticipatory methods which can prompt a focussed civic discourse around significant developments and provide a wider range of people with the information they need to contribute to conversations about the future of AI.

We present a methodology for identifying early warning signs of potentially transformative impacts of AI and discuss how these can feed into more anticipatory and democratic governance processes. We call these early warning signs 'canaries' based on the practice of using canaries to provide early warnings of unsafe air pollution in coal mines in the industrial revolution. Others before us have used this term in the context of AI to stress the importance of early warning

* Corresponding author.
E-mail address: carla.cremer@philosophy.ox.ac.uk

signs [13], [14], but this is the first attempt to outline in detail how such 'artificial canaries' might be identified and used.

Our methodology is a prototype but we believe it provides an important first step towards assessing and then trialling the feasibility of identifying canaries. We first present the approach and then illustrate it on two high-level examples, in which we identify preliminary warning signs of AI applications that could undermine democracy, and warning signs of progress towards high-level machine intelligence (HLMI). We explain why early warning signs are needed by drawing on the literature of participatory technology assessments, and we discuss the advantages and practical challenges of this method in the hope of preparing future research that might attempt to put this method into practise. Our theoretical exploration of a method to identify early warning signs of transformative applications provides a foundation towards more anticipatory, accountable and democratic governance of AI in practice.

II. Related Work

We rely on two main bodies of work. Our methodology for identifying canaries relies on the literature on *forecasting and monitoring AI*. Our suggestions for how canaries might be used once identified build on work on *participatory technology assessments*, which stresses a more inclusive approach to technology governance. While substantial research exists in both these areas, we believe this is the first piece of work that shows how they could feed into each other.

A. AI Forecasting and Monitoring

Over the past decade, an increasing number of studies have attempted to forecast AI progress. They commonly use expert elicitations to generate probabilistic estimates for when different AI advances and milestones will be achieved [10], [15]–[17]. For example, [16] ask experts about when specific milestones in AI will be achieved, including passing the Turing Test or passing third grade. Both [15] and [10] ask experts to predict the arrival of high-level machine intelligence (HLMI), which the latter define as when "unaided machines can accomplish every task better and more cheaply than human workers".

However, we should be cautious about giving results from these surveys too much weight. These studies have several limitations, including the fact that the questions asked are often ambiguous, that expertise is narrowly defined, and that respondents do not receive training in quantitative forecasting [11], [18]. Experts disagree substantially about when crucial capabilities will be achieved [10], but these surveys cannot tell us who (if anyone) is more accurate in their predictions.

Issues of accuracy and reliability aside, forecasts focused solely on timelines for specific events are limited in how much they can inform our decisions about AI today. While it is interesting to know how much experts disagree on AI progress via these probabilistic estimates, they cannot tell us why experts disagree or what would change their minds. Surveys tell us little about what early warning signs to look out for or where we should place our focus today to shape the future development and impact of AI.

At the same time, several projects, e.g. [19]–[22], have begun to track and measure progress in AI. 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 more important than others for the preparation of transformative applications. Time and attention for tracking progress is limited and it would be helpful if we were able to prioritise and monitor those research areas that are most relevant to mitigating risks.

Recognising some of the limitations of existing work, [23] aims for a more holistic approach to AI forecasting. This framework emphasises the use of the Delphi technique [24] to aggregate different perspectives of a group of experts, and cognitive mapping methods to study how different milestones relate to one another, rather than to simply forecast milestones in isolation. We agree that such methods might address some limitations of previous work in both AI forecasting and monitoring. AI forecasting has focused on timelines for particularly extreme events, but these timelines are subject to enormous uncertainty and do not indicate near-term warning signs. AI measurement initiatives have the opposite limitation: they focus on near-term progress, but with little systematic reflection on which avenues of progress are, from a governance perspective, more important to monitor than others. What is needed 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.

B. Participatory Technology Assessments

Presently, the impacts of AI are largely shaped by a small group of powerful people with a narrow perspective which can be at odds with public interest [25]. Only a few powerful actors, such as governments, defence agencies, and firms the size of Google or Amazon, have the resources to conduct ambitious research projects. Democratic control over these research projects is limited. Governments retain discretion over what gets regulated, large technology firms can distort and avoid policies via intensive lobbying [26] and defence agencies may classify ongoing research.

Recognising these problems, a number of initiatives over the past few years have emphasised the need for wider participation in the development and governance of AI [27]–[29]. In considering how best to achieve this, it is helpful to look to the field of science and technology studies (STS) which has long considered the value of democratising research progress [30], [31]. Several publications refer to the 'participatory turn' [32] in STS and an increasing interest in the role of the non-expert in technology development and assessment [27]. More recently, in the spirit of "democratic experimentation" [33], various methods for civic participation have been developed and trialled, including deliberative polls, citizen juries and scenario exercises [33].

With a widening conception of expertise, a large body of research on "participatory technology assessment" (PTA) has emerged, aiming to examine how we might increase civic participation in how technology is developed, assessed and rolled out. We cannot summarise this wide-ranging and complex body of work fully here. But we point towards some relevant pieces for interested readers to begin with. [34] and [35] present a typology of the methods and goals of participating, which now come in many forms. This means that assessments of the success of PTAs are challenging [33] and ongoing because different studies evaluate different PTA processes against different goals [34]. Yet while scholars recognise remaining limitations of PTAs [31], several arguments for their advantages have been brought forward, ranging from citizen agency to consensus identification and justice. There are good reasons to believe that non-experts possess relevant end-user expertise. They often quickly develop the relevant subject-matter understanding to contribute meaningfully, leading to better epistemic outcomes due to a greater diversity of views which result in a cancellation of errors [36], [37]. To assess the performance of PTAs scholars draw from case studies and identify best practices [38]–[40].

There is an important difference between truly participatory, democratically minded, technology assessments, and consultations that use the public to help legitimise a preconceived technology [41]. The question of how to make PTAs count in established representational

democracies is an ongoing challenge to the field [31], [33]. But [42], who present a recent example of collective technology policy-making, show that success and impact with PTAs is possible. [40] draw from 38 international case studies to extract best practices, building on [38], who showcase great diversity of possible ways in which to draw on the public. Comparing different approaches is difficult, but has been done [39], [43]. [41] present a conceptual framework with which to design and assess PTAs, [44] compares online versus offline methodologies and in [35] we find a typology of various design choices for public engagement mechanisms. See also [45] for a helpful discussion on how to determine the diversity of participants, [46] on what counts as expertise in foresight and [30], [32], [47] for challenges to be aware of in implementing PTAs.

Many before us have noted that we need wider participation in the development and governance of AI, including by calling for the use of PTAs in designing algorithms [48], [49]. We see a need to go beyond greater participation in addressing existing problems with algorithms and propose that wider participation should also be considered in conversations about future AI impacts.

Experts and citizens each have a role to play in ensuring that AI governance is informed by and inclusive of a wide range of knowledge, concerns and perspectives. However, the question of how best to marry expert foresight and citizen engagement is a challenging one. While a full answer to this question is beyond the scope of this paper, what we do offer is a first step: a proposal for how expert elicitation can be used to identify important warnings which can later be used to facilitate timely democratic debate. For such debates to be useful, we first need an idea of which developments on the horizon can be meaningfully assessed and influenced, for which it makes sense to draw on public expertise and limited attention. This is precisely what our method aims to provide.

III. Identifying Early Warning Signs

We believe that identifying canaries for transformative AI is a tractable problem and worth investing research effort in today. Engineering and cognitive development present a proof of principle: 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 [50]. AI progress so far has also 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 [51]. By mapping out the dependencies between different capabilities 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.

The proposed methodology is intended to be highly adaptable and can be used to identify canaries for a number of important potentially transformative events, such as foundational research breakthroughs or the automation of tasks that affect a wide range of jobs. Many types of indicators could be of interest and classed as canaries, including: algorithmic innovation that supports 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).

Given an event for which we wish to identify canaries, our methodology has three essential 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. See Fig. 1 for an illustration. We here deliberately refrain from describing the method with too much specificity, because we want to stress the flexibility of our approach, and recognise that there is currently no one-fits-all approach in forecasting. The method will require adaptation to the particular transformative event in question, but each step of this method is suited for such specifications. We outline example adaptations of the method to particular cases.

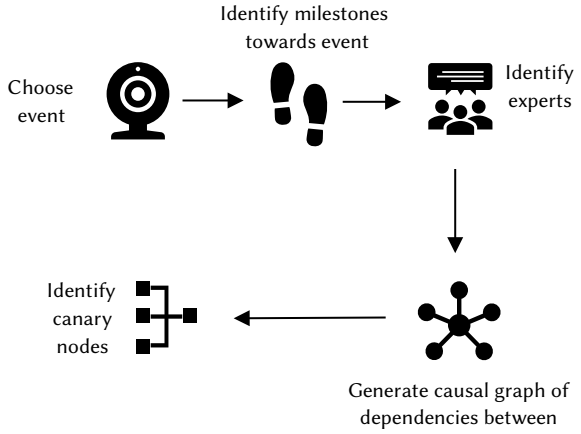

Fig. 1. Illustration of methodological steps to identify canaries of AI progress.

A. Identifying Milestones Via Expert Elicitation

The first step of our methodology involves using traditional approaches in expert elicitation to identify milestones that may be relevant to the transformative event in question. Which experts are selected is crucial to the outcome and reliability of studies in AI forecasting. There are unavoidable limitations of using any form of subjective judgement in forecasting, but these limitations can be minimised by carefully thinking through the group selection. Both the direct expertise of individuals, and how they contribute to the diversity of the overall group, must be considered. See [46] for a discussion of who counts as an expert in forecasting.

Researchers should decide in advance what kinds of expertise are most relevant and must be combined to study the milestones that relate to the transformative event. Milestones might include technical limitations of current methods (e.g. adversarial attacks) and informed speculation about future capabilities (e.g. common sense) that may be important prerequisites to the transformative event. Consulting across a wide range of academic disciplines to order such diverse milestones is important. For example, a cohort of experts identifying and ordering milestones towards HLMI should include not only experts in machine learning and computer science but also cognitive scientists, philosophers, developmental psychologists, evolutionary biologists, or animal cognition experts. Such a group combines expertise on current capabilities in AI, with expertise on key pillars of cognitive development and the order in which cognitive faculties develop in animals. Groups which are diverse (on multiple dimensions) are expected to produce better epistemic outcomes [37], [52].

We encourage the careful design and phrasing of questions to enable participants to make use of their expertise, but refrain from demanding answers that lie outside their area of expertise. For example, asking machine learning researchers directly for milestones towards HLMI does not draw on their expertise. But asking machine learning researchers about the limitations of the methods they use every day; or asking psychologists what human capacities they see lacking in machines today, draws directly on their day-to-day experience. Perceived limitations can be then be transformed into milestones.

There are several different methods available for expert elicitation including surveys, interviews, workshops and focus groups, each with advantages and disadvantages. Interviews provide greater opportunity to tailor questions to the specific expert, but can be time-intensive compared to surveys and reduce the sample size 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.

B. Mapping Causal Relations Between Milestones

The second step of our methodology involves convening experts to identify causal relations between identified milestones: that is, how milestones may underpin, depend on, or affect progress towards other milestones. Experts should be guided in generating directed causal graphs, a type of cognitive map that elicits a person's perceived causal relations between components. Causal graphs use arrows to represent perceived causal relations between nodes, which in this case are milestones [53].

This process primarily focuses on finding out whether or not a relationship exists at all; how precisely this relationship is specified can be adapted to the goals of the study. An arrow from A to B at minimum indicates that progress on A will allow for further progress on B. But this relationship can also be made more precise: in some cases indicating that progress on AI is *necessary* for progress on B, for example. The relationship between nodes may be either linear or non-linear; again this can be specified more precisely if needed or known.

Constructing and debating causal graphs can "help groups to convert tacit knowledge into explicit knowledge" [53]. Causal graphs are used as decision support for individuals or groups, and are often used to solve problems in policy and management involving complex relationships between components in a system by tapping into experts' mental models and intuitions. We therefore suggest that causal graphs are particularly well-suited to eliciting experts' models and assumptions about the relationship between different milestones in AI development.

As a method, causal graphs are highly flexible and can be adapted to the preferred level of detail for a given study: they can be varied in complexity and can be analysed both quantitatively and qualitatively [54], [55]. We neither exclude nor favour quantitative approaches here, due to the complexity and uncertainty of the questions around transformative events. Particularly for very high-level questions, quantitative approaches might not offer much advantage and might communicate a false sense of certainty. In narrower domains where there is more existing evidence, however, quantitative approaches may help to represent differences in the strength of relationships between milestones.

[56] notes that there are no ready-made designs that will fit all studies: design and analysis of causal mapping procedures must be matched to a clear theoretical context and the goal of the study. We highlight a number of different design choices which can be used to adapt the process. As more studies use causal graphs in expert elicitations about AI developments, we can learn from the success of different design choices over time and identify best practices.

[53] stress that interviews or collective brainstorming are the most accepted method for generating the data upon which to analyse causal relations. [57] list heuristics on how to manage the procedure of combining graphs by different participants, or see [58] for a discussion on evaluating different options presented by experts. [59] suggest visual, interactive tools to aid the process. [56] and [60] discuss approaches to analysing graphs and extracting the emergent properties, significant 'core' nodes as well as hierarchical clusters. Core or "potent" nodes are those that relate to many clusters in the graphs and thus have implications for connected nodes. In our proposed methodology, such potent nodes play a central role in pointing to canary milestones. For more detail on the many options on how to generate, analyse and use causal graphs we refer the reader to the volume of [57], or reviews such as [53], [59]. See [55] for an example of applying cognitive mapping to expert views on UK public policies; and [61] for group problem solving with causal graphs.

We propose that identified experts be given instruction in generating either an individual causal graph, after which a mediated discussion between experts generates a shared graph; or that the groups of experts as a whole generates the causal graph via argumentation, visualisations and voting procedures if necessary. As [62] emphasises, any group of experts will have both shared and conflicting assumptions, which causal graphs aim to integrate in a way that approaches greater accuracy than that contained in any single expert viewpoint. The researchers are free to add as much detail to the final maps as required or desired. Each node can be broken into subcomponents or justified with extensive literature reviews.

C. Identifying Canaries

Finally, the resulting causal graphs can be used to identify nodes of particular relevance for progress towards the transformative event in question. This can be a node with a high number of outgoing arrows, i.e. milestones which unlock many others that are prerequisites for the event in question. It can also be a node which functions as a bottleneck - a single dependency node that restricts access to a subsequent highly significant milestone. See Fig. 2 for an illustration. Progress on these milestones can thus represent a 'canary', indicating that further advances in subsequent milestones will become possible and more likely. These canaries can act as early warning signs for potentially rapid and discontinuous progress, or may signal that applications are becoming ready for deployment. Experts identify nodes which unlock or provide a bottleneck for a significant number of other nodes (some amount of discretion from the experts/conveners will be needed to determine what counts as 'significant').

Of course, in some cases generating these causal graphs and using them to identify canaries may be as complicated as a full scientific research project. The difficulty of estimating causal relationships between future technological advances must not be underestimated. However, we believe it to be the case that each individual researcher already does this to some extent, when they chose to prioritise a research project, idea or method over another within a research paradigm. Scientists also debate the most fruitful and promising research avenues and arguably place bets on implicit maps of milestones as they pick a research agenda. The idea is not to generate maps that provide a perfectly accurate indication of warning signs, but to use the wisdom of crowds to make implicit assumptions explicit, creating the best possible estimate of which milestones may provide important indications of future transformative progress.

IV. Using Early Warning Signs

Once identified, canary milestones can immediately help to focus existing efforts in forecasting and anticipatory governance. Given limited resources, early warning signs can direct governance attention to areas of AI progress which are soon likely to impact society and which can be influenced now. For example, if progress in a specific area of NLP (e.g. sentiment analysis) serves as a warning sign for the deployment of more engaging social bots to manipulate voters, policymakers and regulators can monitor or regulate access and research on this research area within NLP.

We can also establish research and policy initiatives to monitor and forecast progress towards canaries. Initiatives might automate

the collection, tracking and flagging of new publications relevant to canary capabilities, and build a database of relevant publications. They might use prediction platforms to enable collective forecasting of progress towards canary capabilities. Foundational research can try to validate hypothesised relationships between milestones or illuminate the societal implications of different milestones.

These forecasting and tracking initiatives can be used to improve policy prioritisation more broadly. For example, if we begin to see substantial progress in an area of AI likely to impact jobs in a particular domain, policymakers can begin preparing for potential unemployment in that sector with greater urgency.

However, we believe the value of early warning signs can go further and support us in democratising the development and deployment of AI. Providing opportunities for participation and control over policy is a fundamental part of living in a democratic society. It may be especially important in the case of AI, since its deployment might indeed transform society across many sectors. If AI applications are to bring benefits across such wide-ranging contexts, AI deployment strategies must consider and be directed by the diverse interests found across those sectors. Interests which are underrepresented at technology firms are otherwise likely to bear the negative impacts.

There is currently an information asymmetry between those developing AI and those impacted by it. Citizens need better information about specific developments and impacts which might affect them. Public attention and funding for deliberation processes is not unlimited, so we need to think carefully about which technologies to direct public attention and funding towards. Identifying early warning signs can help address this issue, by focusing the attention of public debate and directing funding towards deliberation practises that centre around technological advancements on the horizon.

We believe early warning signs may be particularly well-suited to feed into participatory technology assessments (PTAs), as introduced earlier. Early warning signs can provide a concrete focal point for citizens and domain experts to collectively discuss concerns. Having identified a specific warning sign, various PTA formats could be suited to consult citizens who are especially likely to be impacted. PTAs come in many forms and a full analysis of which design is best suited to assessing particular AI applications is beyond the scope of this article. But the options are plenty and PTAs show much potential (see section 2). For example, Taiwan has had remarkable success and engagement with an open consultation of citizens on complex technology policy questions [42]. An impact assessment of PTA is not a simple task, but we hypothesise that carefully designed, inclusive PTAs would present a great improvement over how AI is currently developed, deployed and governed. Our suggestion is not limited to governmental bodies. PTAs or other deliberative processes can be run by research groups and private institutions such as AI labs, technology companies and think tanks who are concerned with ensuring AI benefits all of humanity.

V. Method Illustrations

We outline two examples of how this methodology could be adapted and implemented: one focused on identifying warning signs of a particular societal impact, the other on warning signs of progress towards particular technical capabilities. Both these examples pertain to high-level, complex questions about the future development and impacts of AI, meaning our discussion can only begin to illustrate what the process of identifying canaries would look like, and what questions such a process might raise. Since the results are only the suggestions of the authors of this paper, we do not show a full implementation of the method whose value lies in letting a group of experts deliberate. As mentioned previously, the work of generating these causal maps will often be a research project of its own, and we will return later to the question of what level of detail and certainty is needed to make the resulting graphs useful.

A. First Illustration: AI Applications in Voter Manipulation

We show how our method could identify warning signs of the kind of algorithmic progress which could improve the effectiveness of, or reduce the cost of, algorithmic election manipulation. The use of algorithms in attempts to manipulate election results incur great risk for the epistemic resilience of democratic countries [63]–[65].

Manipulations of public opinion by national and commercial actors are not a new phenomenon. [66] details the history of how newly emerging technologies are often used for this purpose. But recent advances in deep learning techniques, as well as the widespread use of social media, have introduced easy and more effective mechanisms for influencing opinions and behaviour. [8] and [67] detail the various ways in which political and commercial actors incur harm to the information ecosystem via the use of algorithms. Manipulators profile voters to identify susceptible targets on social media, distribute micro-targeted advertising, spread misinformation about policies of the opposing candidate and try to convince unwanted voters not to vote. Automation plays a large role in influencing online public discourse. Publications like [68], [69] note that manipulators use both human-run accounts and bots [70] or a combination of the two [71]. Misinformation [72] and targeted messaging [73] can have transformative implications for the resilience of democracies and very possibility of collective action [74], [75].

Despite attempts by national and sub-national actors to apply algorithms to influence elections, their impact so far has been contested [76]. Yet, foreign actors and national political campaigns will continue to have incentives and substantial resources to invest in such campaigns, suggesting their efforts are unlikely to wane in future. We may thus inquire what kinds of technological progress would increase the risk that elections can be successfully manipulated. We can begin this inquiry by identifying what technological barriers currently prevent full-scale election manipulation.

We would identify those technological limitations by drawing on the expertise of actors who are directly affected by these bottlenecks. Those might be managers of online political campaigns and foreign consulting firms (as described in [8]), who specialise in influencing public opinion via social media, or governmental organisations across the world who comment on posts, target individual influencers and operate fake accounts to uphold and spread particular beliefs. People who run such political cyber campaigns have knowledge of what technological bottlenecks still constrain their influence on voter decisions. We recommend running a series of interviews to collect a list of limitations.

This list might include, for example, that the natural language functionality of social bots is a major bottleneck for effective online influence (for the plausibility of this being an important technical factor see [8]). Targeted users often disengage from a chat conversation after detecting that they are exchanging messages with social bots. Low retention time is presumably a bottleneck for further manipulation, which suggests that improvements in natural language processing (NLP) would significantly reduce the cost of manipulation as social bots become more effective.

We will assume, for the purpose of this illustration that NLP were to be identified as a key bottleneck. We would then seek to gather experts (e.g. in a workshop) who can identify and map milestones (or current limitations) in NLP likely to be relevant to improving the functionality of social bots. This will include machine learning experts who specialise in NLP and understand the technical barriers to developing more convincing social bots; as well as experts in developmental

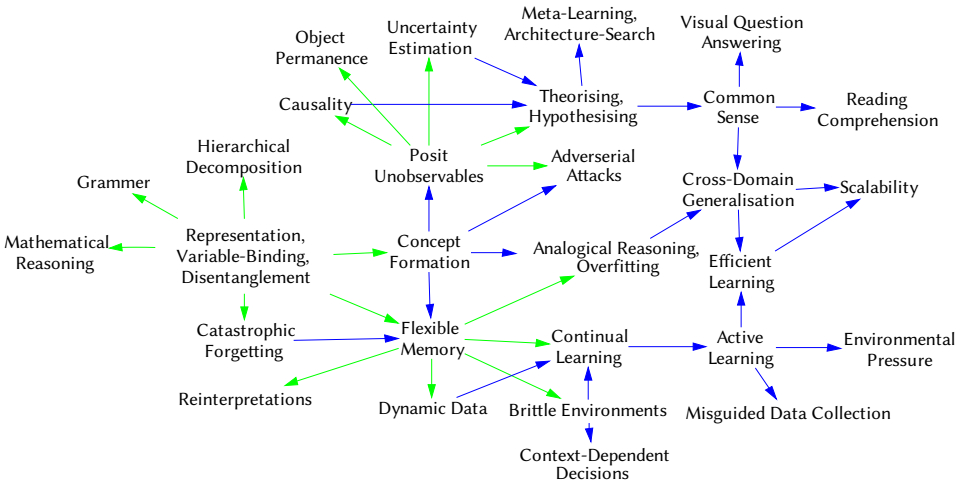

Fig. 2. Cognitive map of dependencies between milestones collected in expert elicitations. Arrows coloured in green signify those milestones that have most outgoing arrows. See appendix for description of each milestone and dependency relations between one 'canary' node and subsequent nodes.

linguistics and evolutionary biology, who can determine suitable benchmarks and the required skills, and who understand the order in which linguistic skills are usually developed in animals.

From these expert elicitation processes we would acquire a list of milestones in NLP which, if achieved, would likely lower the cost and increase the effectiveness of online manipulation. Experts would then order milestones into a causal graph of dependencies. Given the interdisciplinary nature of the question at hand, we suggest in this case that the graph should be directly developed by the whole group. A mediated discussion in a workshop context can help to draw out different connections between milestones and the reasoning behind them, ensuring participants do not make judgements outside their range of expertise. A voting procedure such as majority voting should be used if no consensus can be reached. In a final step, experts can highlight milestone nodes in the final graph which are either marked by many outgoing nodes or are bottlenecks for a series of subsequent nodes that are not accessed by an alternative pathway. These (e.g. sentiment analysis) are our canaries: areas of progress which serve as a warning sign of NLP being applied more effectively in voter manipulation.

Having looked at how this methodology can be used to identify warning signs of a specific societal impact, we next illustrate a different application of the method in which we aim to identify warning signs of a research breakthrough.

B. Second Illustration: High-level Machine intelligence

We use this second example to illustrate in more detail what the process of developing a causal map might look like once initial milestones have been identified, and how canary capabilities can be identified from the map.

We define high-level machine intelligence (HLMI) 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. We choose to focus on HLMI since it is a milestone which has been the focus of previous forecasting studies [10], [15], and which, despite the ambiguity and uncertain nature of the concepts, is interesting to attempt to examine, because it is likely to precipitate widely transformative societal impacts.

To trial this method, we used interview results from [11]. 25 experts from a diverse set of disciplines (including computer science, cognitive science and neuroscience) were interviewed and asked what they believed to be the main limitations preventing current machine learning methods from achieving the capabilities of HLMI. These limitations can be translated into 'milestones': capabilities experts believe machine learning methods need to achieve on the path to HLMI, i.e. the output of step 1 of our methodology.

Having identified key milestones, step 2 of our methodology involves exploring dependencies between them using causal graphs. We use the software VenSim to illustrate hypothesised relationships between milestones (see Fig. 2). For example, we hypothesise that the ability to formulate, comprehend and manipulate abstract concepts may be an important prerequisite to the ability to account for unobservable phenomena, which is in turn important for reasoning about causality. This map of causal relations and dependencies was constructed by the authors alone, and is therefore far from definitive, but provides a useful illustration of the kind of output this methodology can produce.

Based on this causal map, we can identify three candidates for canary capabilities:

Representations that allow variable-binding and disentanglement: 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 memory and knowledge representations. We hypothesise that this ability would unlock many others, including the ability to learn from dynamic data, to learn in a continual fashion, and to update old interpretations of data as new information is acquired.

Positing unobservables: the ability to recognise and use unobservable concepts that are not represented in the visual features of a scene, including numerosity or intentionality.

We might tentatively suggest that these are important capabilities to track progress on from the perspective of anticipating HLMI.

VI. Discussion and Future Directions

As the two illustrative examples show, there are many complexities and challenges involved in putting this method into practice. One particular challenge is that there is likely to be substantial uncertainty in the causal graphs developed. This uncertainty can come in many forms.

Milestones that are not well understood are likely to be composed of several sub-milestones. As more research is produced, the graph will be in need of revision. Some such revisions may include the addition of connections between milestones that were previously not foreseen,

which in turn might alter the number of outgoing connections from nodes and turn them into potent nodes, i.e. 'canaries'.

The process of involving a diversity of experts in a multi-stage, collaborative process is designed to reduce this uncertainty by allowing for the identification of nodes and relationships that are widely agreed upon and so more likely to be robust. However, considerable uncertainty will inevitably remain due to the nature of forecasting. The higher the level of abstraction and ambiguity in the events studied (like events such as HLMI, which we use for our illustration) the greater the uncertainty inherent in the map and the less reliable the forecasts will likely be. It will be important to find ways to acknowledge and represent this uncertainty in the maps developed and conclusions drawn from them. This might include marking uncertainties in the graph and taking this into account when identifying and communicating 'canary' nodes.

Given the uncertainty inherent in forecasting, we must consider what kinds of inevitable misjudgements are most important to try to avoid. A precautionary perspective would suggest it is better to slightly overspend resources on monitoring canaries that turn out to be false positives, rather than to miss an opportunity to anticipate significant technological impacts. This suggests we may want to set a low threshold for what should be considered a 'canary' in the final stage of the method.

The uncertainty raises an important question: will it on average be better to have an imperfect, uncertain mapping of milestones rather than none at all? There is some chance that incorrect estimates of 'canaries' could be harmful. An incorrect mapping could focus undue attention on some avenue of AI progress, waste resources or distract from more important issues.

Our view is that it is nonetheless preferable to attempt a prioritisation. The realistic alternative is that anticipatory governance is not attempted or informed by scholars' individual estimates in an ad-hoc manner, which we should expect to be incorrect more often than our collective and structured expert elicitation. How accurate our method is can only be studied by trialling it and tracking its predictions as AI research progresses to confirm or refute the forecasts.

Future studies are likely to face several trade-offs in managing the uncertainty. For example, a large and cognitively diverse expert group may be better placed to develop robust maps eventually, but this may be a much more challenging process than doing it with a smaller, less diverse group -- making the latter a tempting choice (see [45] for a discussion of this trade-off). The study of broad and high-level questions (such as when we might attain HLMI or automate a large percentage of jobs) may be more societally relevant or intellectually motivating, but narrower studies focused on nearer-term, well-defined applications or impacts may be easier to reach certainty on.

A further risk is that this method, intended to identify warning signs so as to give time to debate transformative applications, may inadvertently speed up progress towards AI capabilities and applications. By fostering expert deliberation and mapping milestones, it is likely that important research projects and goals are highlighted and the field's research roadmap is improved. This means our method must be used with caution.

However, we do not believe this is a reason to abandon the approach, since these concerns must be balanced against the benefits of being able to deliberate upon and shape the impacts of AI in advance. In particular, we believe that the process of distilling information from experts in a way that can be communicated to wider society, including those currently underrepresented in debates about the future of AI, is likely to have many more benefits than costs.

The idea that we can identify 'warning signs' for progress assumes that there will be some time lag between progress on milestones, during which anticipatory governance work can take place. Of course, the extent to which this is possible will vary, and in some cases, unlocking a 'canary' capability could lead to very rapid progress on subsequent milestones. Future work could consider how to incorporate assessment of timescales into the causal graphs developed, so that it is easier to identify canaries which warn of future progress while allowing time to prepare.

Future work should also critically consider what constitutes relevant 'expertise' for the task of identifying canaries, and further explore ways to effectively integrate expert knowledge with the values and perspectives of diverse publics. Our method finds a role for the expert situated in a larger democratic process of anticipating and regulating emerging technologies. Expert judgement can thereby be beneficial to wider participation. However, processes that allow more interaction between experts and citizens could be even more effective. One limitation of the method presented in this paper is that it requires one to have already identified a particular transformative event of concern, but does not provide guidance on how to identify and prioritise between events. It may be valuable to consider how citizens that are impacted by technology can play a role in identifying initial areas of concern, which can then feed into this process of expert elicitation to address the concerns.

VII. Conclusion

We have presented a flexible method for identifying early warning signs, or 'canaries' in AI progress. Once identified, these canaries can provide focal points for anticipatory governance efforts, and can form the basis for meaningful participatory processes enabling citizens to steer AI developments and their impacts. Future work must now test this method by putting it into practice, which will more clearly reveal both benefits and limitations. Our artificial canaries offer a chance for forward-looking, democratic assessments of transformative technologies.

Appendix

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.

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 Fig. 2 (see Fig. 3):

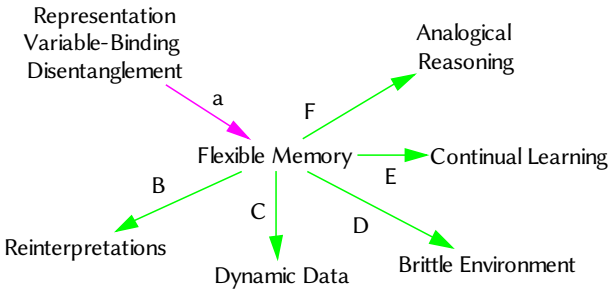

Fig. 3. Extract of Fig. 2, showing one candidate canary capability.

- (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.

Acknowledgements

We thank reviewers for their particularly detailed comments and engagement with this paper, the scholars at the Leverhulme Centre for the Future of Intelligence for fruitful discussions after our presentation,

TABLE I. Limitations of Deep Learning As Perceived and Named by Experts Found in [11]

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.

as well as the attendees of the workshop Evaluating Progress in AI at the European Conference on AI (Aug 2020) for recognizing the potential of this work. We particularly thank Carolyn Ashurst and Luke Kemp for their efforts and commentary on our drafts.

References

[1] K. Crawford *et al.*, 'AI Now Report 2019', *AI 2019 Report*, p. 100, 2019.

[2] S. Russell, *Human Compatible*. Viking Press, 2019.

[3] C. Cath, S. Wachter, B. Mittelstadt, M. Taddeo, and L. Floridi, 'Artificial Intelligence and the "Good Society": the US, EU, and UK approach', *Sci. Eng. Ethics*, vol. 24, no. 2, pp. 505–528, Apr. 2018, doi: 10.1007/s11948-017-9901-7.

[4] 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', *2019*, p. 59.

[5] Y. K. Dwivedi *et al.*, 'Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy', *Int. J. Inf. Manag.*, p. 101994, Aug. 2019, doi: 10.1016/j.ijinfomgt.2019.08.002.

[6] R. Gruetzemacher and J. Whittlestone, 'The Transformative Potential of Artificial Intelligence', *ArXiv191200747 Cs*, Sep. 2020, Accessed: Jan. 09, 2021. [Online]. Available: http://arxiv.org/abs/1912.00747.

[7] M. Brundage *et al.*, 'The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation', *ArXiv180207228 Cs*, Feb. 2018, Accessed: Jan. 15, 2021. [Online]. Available: http://arxiv.org/abs/1802.07228.

[8] P. Howard, *Lie Machines, How to Save Democracy from Troll Armies, Deceitful Robots, Junk News Operations, and Political Operatives*. Yale: Yale University Press, 2020.

[9] C. B. Frey and M. A. Osborne, 'The future of employment: How susceptible are jobs to computerisation?', *Technol. Forecast. Soc. Change*, vol. 114, pp. 254–280, Jan. 2017, doi: 10.1016/j.techfore.2016.08.019.

[10] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, and O. Evans, 'Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts', *J. Artif. Intell. Res.*, vol. 62, pp. 729–754, Jul. 2018, doi: 10.1613/jair.1.11222.

[11] C. Z. Cremer, 'Deep Limitations? Examining Expert Disagreement over Deep Learning', *Prog. Artif. Intell. Springer*, to be published 2021.

[12] D. Collingridge, *The social control of technology*. London: Frances Pinter, 1980.

[13] O. Etzioni, 'How to know if artificial intelligence is about to destroy civilization', *MIT Technology Review*. https://www.technologyreview.com/s/615264/artificial-intelligence-destroy-civilization-canaries-robot-overlords-take-over-world-ai/ (accessed Mar. 12, 2020).

[14] A. Dafoe, 'The academics preparing for the possibility that AI will destabilise global politics', *80,000 Hours*, 2018. https://80000hours.org/podcast/episodes/allan-dafoe-politics-of-ai/ (accessed Jan. 15, 2021).

[15] V. C. Müller and N. Bostrom, 'Future Progress in Artificial Intelligence: A Survey of Expert Opinion', in *Fundamental Issues of Artificial Intelligence*, V. C. Müller, Ed. Cham: Springer International Publishing, 2016, pp. 555–572.

[16] S. D. Baum, B. Goertzel, and T. G. Goertzel, 'How long until human-level AI? Results from an expert assessment', *Technol. Forecast. Soc. Change*, vol. 78, no. 1, pp. 185–195, Jan. 2011, doi: 10.1016/j.techfore.2010.09.006.

[17] S. Beard, T. Rowe, and J. Fox, 'An analysis and evaluation of methods currently used to quantify the likelihood of existential hazards', *Futures*, vol. 115, p. 102469, Jan. 2020, doi: 10.1016/j.futures.2019.102469.

[18] P. E. Tetlock and D. Gardner, *Superforecasting: the art and science of prediction*, First edition. New York: Crown Publishers, 2015.

[19] N. Benaich and I. Hogarth, 'State of AI Report 2020', 2020. https://www.stateof.ai/ (accessed Jan. 15, 2021).

[20] P. Eckersley and Y. Nasser, 'AI Progress Measurement', *Electronic Frontier Foundation*, Jun. 12, 2017. https://www.eff.org/ai/metrics (accessed Jan. 15, 2021).

[21] 'Papers with Code', Available at: https://paperswithcode.com (accessed Feb. 08, 2021).

[22] R. Perrault *et al.*, 'The AI Index 2019 Annual Report', *AI Index Steer. Comm. Hum.-Centered AI Inst. Stanf. Univ. Stanf. CA*, 2019.

[23] Gruetzemacher, 'A Holistic Framework for Forecasting Transformative AI', *Big Data Cogn. Comput.*, vol. 3, no. 3, p. 35, Jun. 2019, doi: 10.3390/bdcc3030035.

[24] H. A. Linstone and M. Turoff, *The delphi method*. Addison-Wesley Reading, MA, 1975.

[25] S. M. West, M. Whittaker, and K. Crawford, 'Discriminating Systems: Gender, Race and Power in AI', AI Now Institute, 2019. [Online]. Available: Retrieved from https://ainowinstitute.org/discriminatingsystems.html.

[26] P. Nemitz and M. Pfeffer, *Prinzip Mensch - Macht, Freiheit und Demokratie im Zeitalter der Künstlichen Intelligenz*. Verlag J.H.W. Dietz Nachf., 2020.

[27] M. Ipsos, 'Public views of Machine Learning: Findings from public research and engagement conducted on behalf of the Royal Society', THe Royal Society, 2017. [Online]. Available: https://royalsociety.org/-/media/policy/projects/machine-learning/publications/public-views-of-machine-learning-ipsos-mori.pdf.

[28] The RSA, 'Artificial Intelligence: Real Public Engagement.', oyal Society for the encouragement of Arts, Manufactures and Commerce, London, 2018.

[29] T. Cohen, J. Stilgoe, and C. Cavoli, 'Reframing the governance of automotive automation: insights from UK stakeholder workshops', *J. Responsible Innov.*, vol. 5, no. 3, pp. 257–279, Sep. 2018, doi: 10.1080/23299460.2018.1495030.

[30] M. Lengwiler, 'Participatory Approaches in Science and Technology: Historical Origins and Current Practices in Critical Perspective', *Sci. Technol. Hum. Values*, vol. 33, no. 2, pp. 186–200, Mar. 2008, doi: 10.1177/0162243907311262.

[31] M. Rask, 'The tragedy of citizen deliberation – two cases of participatory technology assessment', *Technol. Anal. Strateg. Manag.*, vol. 25, no. 1, pp. 39–55, Jan. 2013, doi: 10.1080/09537325.2012.751012.

[32] J. Chilvers, 'Deliberating Competence: Theoretical and Practitioner Perspectives on Effective Participatory Appraisal Practice', *Sci. Technol. Hum. Values*, vol. 33, no. 2, pp. 155–185, Mar. 2008, doi: 10.1177/0162243907307594.

[33] G. Abels, 'Participatory Technology Assessment And The "Institutional Void": Investigating Democratic Theory And Representative Politics" published on 01 Jan 2010 by Brill.', in *Democratic Transgressions of Law*, vol. 112, Brill, 2010, pp. 237–268.

[34] P. Biegelbauer and A. Loeber, 'The Challenge of Citizen Participation to Democracy', *Inst. Für Höhere Stud. - Inst. Adv. Stud. IHS*, p. 46, 2010.

[35] G. Rowe and L. J. Frewer, 'A Typology of Public Engagement Mechanisms', *Sci. Technol. Hum. Values*, vol. 30, no. 2, pp. 251–290, Apr. 2005, doi: 10.1177/0162243904271724.

[36] L. Hong and S. E. Page, 'Groups of diverse problem solvers can outperform groups of high-ability problem solvers', *Proc. Natl. Acad. Sci.*, vol. 101, no. 46, pp. 16385–16389, Nov. 2004, doi: 10.1073/pnas.0403723101.

[37] H. Landemore, *Democratic Reason*. Princeton: Princeton University Press, 2017.

[38] S. Joss and S. Bellucci, *Participatory Technology Assessment: European Perspectives*. London: Center for the Study of Democracy, 2002.

[39] Y. Zhao, C. Fautz, L. Hennen, K. R. Srinivas, and Q. Li, 'Public Engagement in the Governance of Science and Technology', in *Science and Technology Governance and Ethics: A Global Perspective from Europe, India and China*, M. Ladikas, S. Chaturvedi, Y. Zhao, and D. Stemerding, Eds. Cham: Springer International Publishing, 2015, pp. 39–51.

[40] M. T. Rask *et al.*, *Public Participation, Science and Society: Tools for Dynamic and Responsible Governance of Research and Innovation*. Routledge - Taylor & Francis Group, 2018.

[41] J. Burgess and J. Chilvers, 'Upping the ante: a conceptual framework for designing and evaluating participatory technology assessments', *Sci. Public Policy*, vol. 33, no. 10, pp. 713–728, Dec. 2006, doi: 10.3152/147154306781778551.

[42] Y. T. Hsiao, S.-Y. Lin, A. Tang, D. Narayanan, and C. Sarahe, 'vTaiwan: An Empirical Study of Open Consultation Process in Taiwan', SocArXiv, preprint, Jul. 2018. doi: 10.31235/osf.io/xyhft.

[43] J. Hansen, 'Operationalising the public in participatory technology assessment: A framework for comparison applied to three cases', *Sci. Public Policy*, vol. 33, no. 8, pp. 571–584, Oct. 2006, doi: 10.3152/147154306781778678.

[44] T.-P. Ertiö, P. Tuominen, and M. Rask, 'Turning Ideas into Proposals: A Case for Blended Participation During the Participatory Budgeting Trial in Helsinki', in *Electronic Participation: ePart 2019*, Jul. 2019, pp. 15–25,

doi: 10.1007/978-3-030-27397-2_2.
[45] M. Rask, 'Foresight — balancing between increasing variety and productive convergence', *Technol. Forecast. Soc. Change - TECHNOL FORECAST SOC CHANGE*, vol. 75, pp. 1157–1175, Oct. 2008, doi: 10.1016/j.techfore.2007.12.002.
[46] S. Mauksch, H. A. von der Gracht, and T. J. Gordon, 'Who is an expert for foresight? A review of identification methods', *Technol. Forecast. Soc. Change*, vol. 154, p. 119982, May 2020, doi: 10.1016/j.techfore.2020.119982.
[47] J. Saldivar, C. Parra, M. Alcaraz, R. Arteta, and L. Cernuzzi, 'Civic Technology for Social Innovation: A Systematic Literature Review', *Comput. Support. Coop. Work CSCW*, vol. 28, no. 1–2, pp. 169–207, Apr. 2019, doi: 10.1007/s10606-018-9311-7.
[48] T. Kariotis and J. Darakhshan, 'Fighting Back Algocracy: The need for new participatory approaches to technology assessment', in *Proceedings of the 16th Participatory Design Conference 2020 - Participation(s) Otherwise - Volume 2*, Manizales Colombia, Jun. 2020, pp. 148–153, doi: 10.1145/3384772.3385151.
[49] M. Whitman, C. Hsiang, and K. Roark, 'Potential for participatory big data ethics and algorithm design: a scoping mapping review', in *Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial - Volume 2*, New York, NY, USA, Aug. 2018, pp. 1–6, doi: 10.1145/3210604.3210644.
[50] C. Buckner and K. Yang, 'Mating dances and the evolution of language: What's the next step?', *Biol. Philos.*, vol. 32, 2017, doi: 10.1007/s10539-017-9605-z.
[51] Y. LeCun, Y. Bengio, and G. Hinton, 'Deep learning', *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
[52] S. E. Page, *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools and Societies*. Princeton: Princeton University Press, 2008.
[53] A. J. Scavarda, T. Bouzdine-Chameeva, S. M. Goldstein, J. M. Hays, and A. V. Hill, 'A Review of the Causal Mapping Practice and Research Literature', in *Abstract number: 002-0256*, Cancun, Mexico, 2004, p. 21.
[54] L. Markíczy, and J. Goldberg, 'A method for eliciting and comparing causal maps', *J. Manag.*, vol. 21, no. 2, pp. 305–333, Jan. 1995, doi: 10.1016/0149-2063(95)90060-8.
[55] C. Eden and F. Ackermann, 'Cognitive mapping expert views for policy analysis in the public sector', *Eur. J. Oper. Res.*, vol. 152, no. 3, pp. 615–630, Feb. 2004, doi: 10.1016/S0377-2217(03)00061-4.
[56] C. Eden, 'ON THE NATURE OF COGNITIVE MAPS', 1992, doi: 10.1111/J.1467-6486.1992.TB00664.X.
[57] F. Ackerman, J. Bryson, and C. Eden, *Visible Thinking, Unlocking Causal Mapping for Practical Business Results*. John Wiley & Sons, 2004.
[58] G. Montibeller and V. Belton, 'Causal maps and the evaluation of decision options—a review', *J. Oper. Res. Soc.*, vol. 57, no. 7, pp. 779–791, Jul. 2006, doi: 10.1057/palgrave.jors.2602214.
[59] A. J. Scavarda, T. Bouzdine-Chameeva, S. M. Goldstein, J. M. Hays, and A. V. Hill, 'A Methodology for Constructing Collective Causal Maps*', *Decis. Sci.*, vol. 37, no. 2, pp. 263–283, May 2006, doi: 10.1111/j.1540-5915.2006.00124.x.
[60] C. Eden, F. Ackermann, and S. Cropper, 'The Analysis of Cause Maps', *J. Manag. Stud.*, vol. 29, no. 3, pp. 309–324, 1992, doi: https://doi.org/10.1111/j.1467-6486.1992.tb00667.x.
[61] F. Ackermann and C. Eden, 'Using Causal Mapping with Group Support Systems to Elicit an Understanding of Failure in Complex Projects: Some Implications for Organizational Research', *Group Decis. Negot.*, vol. 14, no. 5, pp. 355–376, Sep. 2005, doi: 10.1007/s10726-005-8917-6.
[62] C. Eden, F. Ackermann, J. Bryson, G. Richardson, D. Andersen, and C. Finn, 'Integrating Modes of Policy Analysis and Strategic Management Practice: Requisite Elements and Dilemmas', p. 13, 2009.
[63] L.-M. Neudert and P. Howard, 'Ready to vote: elections, technology and political campaigning in the United Kingdom', Oxford Technology and Elections Commission, Report, Oct. 2019. Accessed: Jan. 11, 2021. [Online]. Available: https://apo.org.au/node/263976.
[64] G. Bolsover and P. Howard, 'Computational Propaganda and Political Big Data: Moving Toward a More Critical Research Agenda', *Big Data*, vol. 5, no. 4, pp. 273–276, Dec. 2017, doi: czz.
[65] M. J. Mazarr, R. Bauer, A. Casey, S. Heintz, and L. J. Matthews, 'The Emerging Risk of Virtual Societal Warfare: Social Manipulation in a Changing Information Environment', Oct. 2019, Accessed: Jan. 14, 2021. [Online]. Available: https://www.rand.org/pubs/research_reports/RR2714.html.
[66] T. Wu, *The Attention Merchants: From the Daily Newspaper to Social Media, How Our Time and Attention is Harvested and Sold*. London: Antlantic Books, 2017.
[67] K. Starbird, 'Disinformation's spread: bots, trolls and all of us', *Nature*, vol. 571, no. 7766, pp. 449–450, Jul. 2019.
[68] R. Gorwa and D. Guilbeault, 'Unpacking the Social Media Bot: A Typology to Guide Research and Policy', *Policy Internet*, vol. 12, no. 2, pp. 225–248, Jun. 2020, doi: 10.1002/poi3.184.
[69] E. Ferrara, 'Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election', Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2995809, Jun. 2017. doi: 10.2139/ssrn.2995809.
[70] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, 'The spread of low-credibility content by social bots', *Nat. Commun.*, vol. 9, no. 1, Art. no. 1, Nov. 2018, doi: 10.1038/s41467-018-06930-7.
[71] P. N. Howard, S. Woolley, and R. Calo, 'Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration', *J. Inf. Technol. Polit.*, vol. 15, no. 2, pp. 81–93, Apr. 2018, doi: 10.1080/19331681.2018.1448735.
[72] M. Chessen, 'The MADCOM Future: How Artificial Intelligence Will Enhance Computational Propaganda, Reprogram Human Culture, and Threaten Democracy… and What can be Done About It.', in *Artificial Intelligence Safety and Security*, Chapman and Hall/CRC Press, 2018, pp. 127–144.
[73] K. Kertysova, 'Artificial Intelligence and Disinformation: How AI Changes the Way Disinformation is Produced, Disseminated, and Can Be Countered', 2018, doi: 10.1163/18750230-02901005.
[74] J. Brainard and P. R. Hunter, 'Misinformation making a disease outbreak worse: outcomes compared for influenza, monkeypox, and norovirus', *SIMULATION*, vol. 96, no. 4, pp. 365–374, Apr. 2020, doi: 10.1177/0037549719885021.
[75] E. Seger, S. Avin, G. Pearson, M. Briers, S. O Heigeartaigh, and H. Bacon, 'Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically-advanced world', Allan Turing Institute, CSER, dstl, 2020. Accessed: Jan. 15, 2021. [Online]. Available: https://www.turing.ac.uk/sites/default/files/2020-10/epistemic-security-report_final.pdf.
[76] K. H. Jamieson, *Cyberwar: How Russian Hackers and Trolls Helped Elect a President: What We Don't, Can't, and Do Know*. Oxford University Press, 2020.

Carla Zoe Cremer

Carla Zoe is a Research Scholar at the Future of Humanity Institute at the University of Oxford and a Research Affiliate at the Centre for the Study of Existential Risk at the University of Cambridge. Her background is in neurobiology, acquired at Ludwig-Maximilian University in Munich and ETH Zurich. She works on comparative cognition, the limitations of deep learning, and on estimating tail-risks of emerging technologies.

Jess Whittlestone

Jess Whittlestone is a Senior Research Associate at Centre for the Study of Existential Risk and the Leverhulme Centre for the Future of Intelligence at the University of Cambridge. She works on various aspects of AI ethics and policy, with a particular focus on what we can do today to ensure AI is safe and beneficial in the long-term. She holds a PhD in Behavioural Science from the University of Warwick and a degree in Mathematics and Philosophy from Oxford University.