Artificial Intelligence's Societal Impacts, Governance, and Ethics: Introduction to the 2019 Summer Institute on AI and Society and its rapid outputs

by: Edward Parson, Alona Fyshe and Dan Lizotte

The works assembled here are the initial outputs of the First International Summer Institute on Artificial Intelligence and Society (SAIS). The Summer Institute was convened from July 21 to 24, 2019 at the Alberta Machine Intelligence Institute (Amii) in Edmonton, in conjunction with the 2019 Deep Learning/Reinforcement Learning Summer School. The Summer Institute was jointly sponsored by the AI Pulse project of the UCLA School of Law (funded by a generous grant from the Open Philanthropy Project) and the Canadian Institute for Advanced Research (CIFAR), and was co-organized by Ted Parson (UCLA School of Law), Alona Fyshe (University of Alberta and Amii), and Dan Lizotte (University of Western Ontario). The Summer Institute brought together a distinguished international group of 80 researchers, professionals, and advanced students from a wide range of disciplines and areas of expertise, for three days of intensive mutual instruction and collaborative work on the societal implications of AI, machine learning, and related technologies. The scope of discussions at the Summer Institute was broad, including all aspects of the societal impacts of AI, alternative approaches to their governance, and associated ethical issues.

Inspired by recent triumphs in machine learning applications, issues of the societal impacts, governance, and ethics of these technologies are seeing a surge of concern, research and policy attention. These rapid linked advances – in multiple linked areas of algorithm development, data and data-handling tools, and hardware-based computational ability – are a leading current concern about technology's potential for profound and disruptive societal transformation.

In part, current concerns about AI reprise familiar themes from other areas of high-stakes technological advance, so the existing body of research on these other technology areas offers insights relevant for AI. A few of these insights are

by: Edward Parson, Alona Fyshe and Dan Lizotte

especially prominent. For example, the rate and character of technological change are shaped not just by scientific knowledge but also by the economic, policy/legal, and social/cultural conditions that determine relevant actors' incentives and opportunities. Societal impacts are not intrinsic to characteristics of technology, but depend strongly on how it is developed, integrated into products and services, and used – and how people adjust their behavior around it: As Kranzberg's first law of technology tells us, "Technology is neither good nor bad; nor is it neutral."[1] Together, the conjunction of rapid technical change and uncertain uses and responses challenge efforts to govern the associated impacts, so governance often merely aims to mitigate the worst impacts after the fact. Even when societal impacts are profound, they tend to emerge gradually in response to repeated adaptations of technology, deployment, and behavior, and are thus difficult to project, assess, or manage in advance.

These broad parallels with prior areas of technological advance and associated societal concerns are real, but there are also reasons to expect that AI may be different, and more serious, in its impacts and implications. What is popularly called "AI" is not one thing, but a cluster of multiple algorithmic methods, some new and some old, which are linked to parallel advances in the scale and management of data, computational capacity, and multiple related application areas. This set of advancing capabilities is diffuse, labile, and hard to define – a particular challenge to governance, since the ability to workably define something is normally a precondition for any legal or regulatory response. AI is also foundational, potentially able to transform multiple other technologies, research fields, and decision areas – to the extent that its impact has been credibly compared to that of electricity or fossil fuels in prior industrial revolutions.

AI's societal impacts thus present deep uncertainties, for good or ill. Expert views of

what it will do, and how fast, span a broad range: from the cumulation of many incremental changes, to existential transformations of human capabilities, prospects, societies, and identities. Even setting aside "singularity" issues – potential general or super-intelligent AI that might threaten (or in some accounts, transcend) human survival and autonomy – multiple mechanisms of impact have been identified by which even continued development of AI short of these landmarks could have transformative societal impacts. Examples include large-scale displacement of human livelihoods, disruption of geopolitical security relationships, transforming (or undermining) collective decision-making processes through democratic governments or other institutions, extreme concentration of wealth and power (perhaps based on new mechanisms of power), and large-scale changes in human capabilities and identities. Even limiting attention to present and near-term developments, there are a host of concerns raised by current AI applications – e.g., safety and security of systems, bias in algorithmic decision-making, threats to privacy, and inscrutability of decisions – some of which may also give early warning signs of coming larger-scale impacts.

Relative to the scale and gravity of potential impacts, present debate on AI and Society presents a seeming paradox. The issue is receiving a flood of attention, with dozens of new programs, a rapid flow of resources, and meetings and conferences seemingly every week. Yet well-founded insights remain scarce on the nature and mechanisms of impacts, effective and feasible means of governing them, and associated ethical issues. There has been relatively little convergence or progress on major questions, which in many cases remain not just unanswered but also subject to wide uncertainty and disagreement, or even not yet clearly posed.[2] Because AI is so labile and weakly defined, studying its impacts has been likened to the ancient Buddhist parable of the blind men and the elephant: each observer feels that part of

the unfamiliar thing that is closest to them, so each thinks they know it; yet their views are all partial and mutually contradictory. As with the elephant, it is possible to approach AI impacts from any discipline or field of inquiry (e.g., corporate law, anthropology, Marxist social history), any area of interest (education, finance, climate change), any political or ethical concern (racial justice, social mobility, privacy, due process), or any prior technological analogy, and find something resonant. Pulled by these centrifugal forces, the debate is thus unhelpfully sub-divided along multiple dimensions and lacks a coherent core.

There is also continued disagreement over where the most important impacts sit in time and scale, yielding a distribution of present attention and concern that is bi-modal. To be a little glib, those whose disciplinary perspectives make them most comfortable with speculative reasoning – often technical AI researchers and philosophers – are attracted to endpoint, singularity-related issues, which lend themselves to elegant, analytically rich theoretical inquiries. Most other researchers, on the other hand, gravitate to current concerns and historical precedents, because their disciplines frown on speculation and favor arguments based on observable (i.e., present or past) data and evidence. These areas of inquiry are both valuable and important, yet they leave disturbingly empty the large middle ground of impacts and challenges lying between these endpoints – where AI might transform people and societies by vastly reconfiguring capabilities, information, and behavior, while still remaining (mostly) under human control.[3] At the same time, while there is a widespread sense that early action is needed to assess and limit risks of severe harmful impacts, there is little knowledge, and even less agreement, on what that action should consist of or how it should be developed.

This description of the range of issues posed by AI and the state of present debate underpin the aims of the Summer Institute. Just as AI and its impacts present a huge

by: Edward Parson, Alona Fyshe and Dan Lizotte

societal challenge, so too does mobilizing existing bodies of experience, knowledge, and methods to effectively inform the assessment and management of its impacts. These challenges will not be surmounted by any single insight, study, or activity. The summer institute aimed to point, tentatively, at a direction of efforts that can advance and expand the debate, establish an early model of the kind of collective engagement needed, and – by seeding cross-disciplinary networks for continued collaborations – contribute to the long-run project of building the needed capacity.

The summer institute pursued this aim in two ways. First, it sought to convene the needed broadly interdisciplinary dialog, with the ability to integrate knowledge and experience from multiple technical, scientific, and humanistic domains, and to resist widespread tendencies to converse mainly within existing disciplinary communities. In seeking this breadth of expertise contributing to the discussions, the summer institute benefited from its co-convening between a program on AI and Society based at a leading law school, and the CIFAR Deep Learning and Reinforcement Learning Summer School – a vehicle for advanced technical AI training with a distinguished international group of faculty and advanced students. Yet even with the right breadth of expertise in the room, making such interdisciplinary interactions productive takes sustained hard work to understand each other, clarify key concepts and methods, and build new conceptual and communication skills. These aims are better advanced by sustained collaborative work on problems of commonly recognized importance than by discussions that lack such common goals, which tend toward superficiality. Second, it is clear that understanding and addressing AI-related impacts is a long-term, even inter-generational project, which must combine mutual instruction with advancing inquiry, aiming to both advance the debate and broaden its participation by engaging more junior and more senior thinkers on collegial terms.

by: Edward Parson, Alona Fyshe and Dan Lizotte

To pursue these aims, the summer institute experimented with a novel two-part structure, with the first part tightly programmed and structured by the organizers and the second part left almost entirely to the collective, bottom-up authority of the group. The first part aimed to provide the essential foundation of common knowledge and concepts to enable people from a wide range of fields and career stages to participate effectively and confidently in the discussions. To this end, the institute opened with a day of short, focused briefings by faculty experts, each covering elements from their expertise they judged essential for anyone to be an informed participant in the debates. These briefings were grouped into four sessions organized by broad subject-matter:

- Recent advances and current technical issues in AI and Machine Learning (briefings by Graham Taylor on Deep Learning, Rich Sutton on Reinforcement Learning, and Dirk Hovy on Natural Language Processing);
- Current issues and controversies in AI societal impacts (briefings by Elizabeth Joh on use of AI in policing and criminal justice; Michael Karlin on military uses of AI; Trooper Sanders on biased data, its implications and potential correctives; and Elana Zeide on use of predictive analytics in education and employment);
- Alternative approaches to governance of AI and its impacts (briefings by Geoffrey Rockwell on the historical trajectory of concerns about automation and proposed responses; Gary Marchant on limits to hard-law approaches, and potential soft-law and international alternatives; Brenda Leong on corporate AI ethics boards and their limitations; and Craig Shank on internal corporate controls and multi-stakeholder governance processes);
- Larger-scale and medium-term issues (briefings by Jason Millar on embedded values in navigation and mobility systems; Evan Selinger on facial recognition and its implications; Osonde Osoba and Casey Bouskill on technology-culture interactions

by: Edward Parson, Alona Fyshe and Dan Lizotte

in AI impacts and governance; and Robert Lempert on large-scale societal implications of alternative approaches to algorithm design).

Following the briefings, the rest of the Summer Institute was dedicated to collaborative work on projects that were not pre-specified, but instead were developed and proposed by individual participants, then selected in real time by all participants choosing which of the proposed projects they wanted to work on. Any participant, regardless of seniority, was invited to propose a workgroup project via a statement posted online and a short oral "pitch" presentation to the group, followed by brief clarifying discussion. In selecting projects, participants were urged to consider a few explicit criteria – that the projects address interesting and important issues related to AI and society, that they not duplicate existing work, and that they offer the prospect of meaningful progress in the limited time available. Otherwise, there was no central control of projects proposed or chosen. The form of proposed projects was completely unconstrained, explicitly including making a start on collaborative research projects, drafting op-eds or other non-specialist publications, developing proposed contributions to policy or governance, developing instructional material, or creating a story or other work of art on the theme of AI and society.

From twelve proposals, the group selected eight highly diverse projects to work on. The resultant eight groups worked intensively over a day and a half, in a process that several participants likened to a hack-a-thon. The analogy is suggestive but only partly accurate, in that that each SI workgroup included a wide range of disciplinary skills and expertise, and each pursued a different project, all generated by participants rather than pre-specified by organizers. The entire group convened briefly in plenary at half-day intervals to hear short reports from each workgroup summarizing what they were doing, what progress they had achieved, what completed output they targeted by the end of the SI, and what help they needed

by: Edward Parson, Alona Fyshe and Dan Lizotte

from the rest of the group. All eight workgroups achieved substantial progress by the end of the summer institute, even within the extremely limited time available. All eight also expressed the intention and developed concrete plans to continue their collaborative work after the Summer Institute – with some continuing that work immediately afterwards.

The contributions published here represent the initial outputs of these eight work groups' collaborative efforts, as achieved during the intensive work period of the summer institute plus a little further polishing over the following few weeks. One consequence of the decentralized, bottom-up model, with each workgroup defining its own project, is that the resultant outputs are too diverse for any single publication or communication outlet to be suitable for them all. Yet in order to have a single vehicle that captures the collective energy and themes of the SI – and moreover, to communicate these while the experience is still fresh in participants' minds – all workgroups agreed to disseminate interim outputs from their work for this fast web publication. This quickly distributed – but explicitly half-baked – publication model was variously likened to theatrical workshopping or rapid prototyping in product development, in addition to hack-a-thons.

This experimental early publication model is very much in line with the exploratory and experimental spirit of the SI, taking the risk of trying different models to advance and broaden the debate. It also, of course, has the unavoidable consequence that these works – while they reflect remarkable achievements in the short time available – are all provisional and not yet fully developed. With some variation among the workgroups, they are presented here with the aim of being starting points for needed discussions, and providing concrete resources, background information, and proposals to move those discussions forward with specificity. They are not completed or polished products.

by: Edward Parson, Alona Fyshe and Dan Lizotte

We provide below a brief synopsis of the aims and outputs of each of the eight workgroups. Each workgroup is continuing to develop its project, aiming for publication in various outlets in line with the groups' diverse aims and intended audiences. As the outlet for each workgroup's completed work is finalized, we will identify it and, as available, add links to the discussions below.

Mobility Systems and Embedded Values

This group used the example of the now-ubiquitous, AI-driven, turn-by-turn navigation systems to illustrate the range of values affected by these systems, whether explicitly or not. They then considered the resultant implications for societal impacts of projected large-scale expansion and integration of these systems, moving from separate navigation apps used by individual drivers, to complete urban mobility systems integrating signaling and multiple types of human-driven and autonomous vehicles, private and public. Navigation apps may at first glance seem prosaic, but the exploration was surprisingly rich. Present implementations of these systems seek to minimize individual drivers' travel time between a given origin and destination, with limited options to tune results to individual preferences such as avoiding freeways. But since their early deployment, a collateral impact of these systems has been increased traffic in residential neighborhoods – an impact well known to the planners who design streets, signals, and signage, but not recognized as implicated in individual navigation systems until large numbers of drivers began taking the same recommended shortcuts through side streets. The group identified several additional values affected by mobility design systems, which will require explicit consideration as the scope and integration of systems increases. In addition to travel time and neighborhood character, these include safety (at the individual level for drivers, pedestrians, and other street users, and collectively); allocation and prioritization of mobility access among types of users (now implemented simply,

through right-of-way for emergency vehicles, HOV or toll lanes, etc., but potentially generalizable in multiple ways with fully integrated systems); and policing strategy and resource allocation, among others – including an unexpected linkage to the important role presently played by traffic fines in some local government budgets. In this initial published collection, the workgroup presents a taste of their discussions in the form of a fictitious press release, announcing the release of a new navigation app that generates routes based on minimizing drivers' cognitive burden.

This group's discussion illustrates a widespread phenomenon related to automation of decision processes. Societal institutions and processes often serve multiple values, only some of which are explicitly articulated as their mission or objective. Just as urban transport systems advance multiple values in addition to efficient mobility, so too do other organizations. A prominent example is provided by military services, in the United States and to different degrees in other countries. While their explicit missions are all broadly related to national defense and security, one of their most important social impacts – almost unrelated to their explicit missions – has long been to provide training and life skills to young people from disadvantaged backgrounds, making these organizations one of the most powerful drivers of social mobility. Many institutions serve such multiple corollary or implicit societal values. Automation or codification of decisions – typically with a single objective function that aligns with the institution's explicit, official mission – can put these other implicated values at risk, either from the automated decisions themselves or from related organizational changes. (In military organizations, the concern arises from the higher level of technical skills and education required of even entry-level recruits in AI-rich environments.) Yet these corollary values are challenging to integrate into explicit algorithmic decision-making – because they are ambiguous, hard to integrate into an objective function that trades them off against core organizational missions,

and potentially contestable – such that they may only flourish while flying under the radar. As Joni Mitchell sang in another context, "You don't know what you've got till it's gone." The loss of corollary, emergent, or ambiguously defined organizational values may be a systematic consequence of automating decisions, which typically requires explicitly codifying what before was ambiguously embedded in organizational practice.[4]

Meaningful Human Control

This group considered the problem of coupled human and algorithmic decision-making in high-stakes settings, using as initial examples the domains of weapons, aviation, and medicine. Noting the definitional ambiguity and difficulty operationalizing widely repeated concepts such as "humans-in-the-loop," their initial ambition was to unpack the meaning of "meaningful human control" (MHC) and identify processes and criteria to operationalize it across these diverse decision domains. But the group adjusted mid-course, recognizing that this was a longer project and that they needed first to engage the prior question of why – and with what conditions and limitations – meaningful human control is judged desirable, or even essential, in such decision contexts. They argue that retaining meaningful human control carries both costs and benefits, and that both the costs and benefits include distinct components, some related to system performance and some to issues of legal and moral responsibility. In general, greater human control may improve system performance by increasing redundancy and adaptability to novel conditions, and may be necessary to ensure moral and legal accountability. Yet it may also degrade performance by requiring uncoupling of complex autonomous systems and increase the risk of human error, carelessness, or other forms of improper human decisions. The group noted that the optimal balancing of these factors, and hence the preferred degree and form of human control, are likely to

vary substantially even among the three decision domains they consider. The group is continuing work on the larger project generating guidelines how to implement the desired degree and form of human control in particular decision types.

AI Without Math

This group began a project to develop non-technical instructional materials on key AI and machine-learning concepts. They recognized that as deployed AI-based products and services continue to expand, many decisions will be required about how to control, explain, and manage these. These decisions will include many by various professionals who not only lack specific training in AI and Machine Learning, but may also lack training in the underlying mathematical and statistical concepts that provide the core of even introductory instruction in AI/ML. In view of this need, the group began development of an online instructional resource that would provide introductory explanations of key AI/ML concepts with no use of formal mathematical notation. As illustrative audiences toward whom to target their explanations, they took journalists and judges. Their short contribution here presents a start on this project and an illustration of their targeted level of explanation, including explanations for four key concepts: rational agents, naïve Bayes classifiers, linear regression, and convolutional neural networks. Their more extensive resource will be an ongoing project, to be available at https://www.aiwithoutmath.com.

Siri Humphrey:[5] *Design Principles for an AI Policy Analyst*

There are many studies underway of the potential for AI tools to take on various functions of government – legislative, executive, judicial, and electoral – asking how the use of AI in specific functions would work, what it would require, with what attendant benefits and risks, and whether (and how) it could align with applicable legal, democratic, and moral principles. This group looked at a previously unexamined piece of this landscape, the potential for AI systems to take over, partly

or wholly, the functions of policy analysts who advise senior officials or political leaders. Starting from recent scholarship that has identified several distinct functions that policy analysts perform, they examined how AI systems – either current ones or reasonably projected extensions – could serve these functions, with what implications for the policy-making process and the multiple public values implicated in policy decisions.

The group argues that AI systems could substantially replace the "synthesis" function of policy analysis: the gathering, curating, and synthesis of publicly available information relevant to an issue or decision. At least initially, use of AI in this role would have to be subject to specific limitations on the tasks delegated, and also subject to review and revision of the resultant briefing notes or other documents before they go to Ministers or other senior decision-makers. The group also argues that repetition of this synthesis and review process, with feedback from both human policy analysts and decision-makers (such as Ministers routinely provide on briefing materials prepared by their human staff) could serve as high-order training for the AI, allowing progressive reduction – although not elimination – of the amount of oversight and input needed from human policy analysts. In contrast to the "synthesis" function, they argue that certain other policy analysis functions depend more strongly on the essentially human interaction between decision-makers and their advisors. This militates against the wholesale replacement of analysis and advising functions by AI systems, suggesting instead a model of "Artificial-intelligence-amplified policy analysis," in which AI systems augment and amplify the skills of human policy analysts.

Assessment Tool for Ethical impacts of AI products

The next two workgroups form a complementary pair, both concerned with the problem of what to do with the multiple sets of AI ethical principles being advanced

to provide guidance for individuals or organizations engaged in AI development and application. These sets of principles pose two widely noted problems. First, the proliferation of large numbers of similar, but not quite identical, lists of principles raise questions about the relationships between them, the normative foundations of any of them, and the basis for adopting any of them over the others.[6] Second, all these principles are stated at high levels of generality and abstraction, so their implied guidance for what to do, or what not to do, in the actual development, design, training, testing, application, and deployment of AI-enabled systems is indirect, non-obvious, and contestable.

In an unplanned piece of serendipity, these two groups approached the same problem from nearly opposite perspectives, one operational and one critical, yielding a rich and instructive counterpoint. This group took an operational, constructive approach rooted in engineering. Boldly (and practically) going where no one has gone before, they reasoned step by step through the process of operationalizing a particular set of ethical principles for any AI-related product or project. They first reduced each principle to a list of specific areas of concern, then to operational questions about observable practices and procedures relevant to each area of concern, and finally to a numerical scoring system for alternative answers to each question. Subject to some remaining ambiguities about appropriate weighting, the resultant component scores can then be aggregated to generate an overall numerical score for conformity of a system or project with the specified principle. The group stresses that such reductive scoring systems are prone to various forms of misinterpretation and misuse – such as imputing false precision or prematurely closing discussions. They also highlight that this heroic, first-cut effort is incomplete. Yet at the same time, they vigorously defend the approach as providing a stimulus, and a concrete starting point, for the discussions of impacts and ethical

implications that are needed in the context of specific projects and systems.

From Shortcut to Sleight-of-Hand: Why the checklist approach in the EU guidelines does not work

This group took as their starting point a different set of ethical principles, the "Ethics Guidelines for Trustworthy Artificial Intelligence" issued by the EU Commission's High-Level Expert Group on Artificial Intelligence in April and June 2019, including an "assessment checklist." This checklist is intended to help technology developers consider ethical issues in their policies and investments, and thus to create more trustworthy AI. In effect, this EU expert group undertook an exercise quite similar to that conducted by the Summer Institute "Tools" group summarized above, except that the EU expert group's exercise is more limited: it consists only of a checklist of yes/no questions (with extensive supporting discussion), and does not pursue a numerical scoring system.

This workgroup conducted a detailed critical assessment of the guidelines and checklist, aiming to assess their implications – and in particular, their limitations – as a tool to guide AI development. They argued that these guidelines are a fair target for such critical scrutiny because of their likely influence and importance, based on their ambition to articulate a broadly applicable standard of care for AI development and their prospect of influencing EU regulatory development – especially given the EU's emerging role as a world leader in this regulatory area.

The group finds the proposed approach problematic in several ways, most of them related to intrinsic limitations of checklists in this context rather than problems specific to this particular checklist. Using the analogy of safety procedures in aviation and space flight, they argue that checklists are an appropriate technology to manage human-factors risks in complex environments whose operations, salient risk

mechanisms, and implicated values are well known, but that these conditions do not apply to development of safe or ethical AI systems. The group argues that many items on the checklist are seriously ambiguous but lack the additional explanation or documentation needed to reduce the ambiguity; and that the checklist thus risks conveying false confidence that needed protections are in place, when the conditions for this to be the case are in fact subtle, context-specific, and evolving over time.

Although the EU expert group's report includes extensive discussions of caveats and limitations, the workgroup finds these insufficient to mitigate the risks they identify, in view of the likely uses of the checklist in real-world, operational settings. They worry that enterprises are likely to treat the checklist either reductively or opportunistically – perhaps delegating responses to their legal teams to seek defensible markers of regulatory compliance or fulfilling some relevant duty of care. Used in such ways, the checklist would fail to stimulate the serious, organization-wide reflection on the concrete requirements of ethical conduct in their setting that should be the aim. Moreover, the group argues, the checklist is unlikely ever to yield a decision not to pursue an otherwise attractive project due to irreducible risks of unacceptable outcomes, when a meaningful and effective ethical filter must be capable – at least occasionally – of generating this outcome. Finally, the group argues that checklists are likely to be proposed or used as safe harbors – by enterprises, or even worse, by regulators, judges, citizen groups, or political leaders – with the resultant risk of reducing the pursuit of ethical AI to empty "ethics-washing" or "ethics theatre."

In contrast to their sharp criticism of the checklist, the group finds the expert group's higher-level "guiding questions" to be of great value, in helping to identify issues and problems that require sustained attention and so to promote an

organizational culture of heightened ethical awareness. But they find the pursuit of simplification and codification embodied in the checklist approach to be premature, promoting misleading, too-optimistic assessments of risks and the subsequent prospect of broad, destructive backlash against the AI and related technologies broadly.

AI and Agency

This group examined the deep, and deeply contested, concept of "agency," as it applies to and is modified by the context of AI development. Working both individually and collectively, they wrote a set of short, provocative essays that approach the concept of agency from multiple disciplinary perspectives, including philosophy, political science, sociology, psychology, economics, computer science, and law. The essays also lay out a set of deep questions and tensions inherent in the concept. They ask how agency is defined; whether humans have it, and if so, whether and how this distinguishes humans from present and prospective AI (and also from other animals); and what are the implications of alternative conceptions and ascriptions of agency – for human behavior, identity, welfare, and social order.

The definitions they consider for agency cluster around two poles, one positive and one negative. At the positive pole, agency is defined by the capacity for goal-directed behavior, and thus identified by observing robust pursuit of a goal in response to obstruction. At the negative pole, agency is defined by not being subject to causal explanation without introducing conceptions of intention or subjectivity. The group notes that conventional conceptions of agency as being unique to humans are increasingly challenged on two fronts: by human inequity in diverse social contexts, and hence wide variation in individual humans' capacities to exercise effective agency; and by scientific advances that suggest both that subjectively perceived agency may be illusory, and that to the extent humans do have agency, so

too may other animals.

Present and projected developments of AI raise the stakes of these inquiries. The increasing complexity of AI performance implies, at a minimum, the lengthening of causal chains connecting behavior to proximate or instrumental goals and thence to higher-order goals, shifting the location of agency and casting doubt on simple claims that people have it but AI's do not or cannot. Yet the connection between this causation-driven notion of agency, and thus the validity of societal ascription of responsibility and deployment of incentives, are obscure, in the context of both human and AI decision-making. Does accountability always pass back to the human designer or creator, no matter how many layers of intermediate goals are generated within an AI? If human behavior is increasingly understood as subject to causation, does this reduce moral problems to correctible, technical ones – and if so, correctible by whom, in terms of both effectiveness and legitimacy? Finally, even if strong human-other or subject-object distinctions in ascribing agency become untenable under further advance of scientific knowledge and AI technology, might agency nevertheless be a useful fiction, a myth that is useful or even necessary to believe – for stable conceptions of human identity, and for effective collective regulation of human behavior?

Can AI be an instrument of transformative social and political progress? The "levelers" group

This group took its inspiration from a strain of political thought early in the industrial revolution, which identified markets and technological innovation as powerful engines of political progress, holding the prospect of large gains in both liberty and equality. Looking forward to the transformative possibilities of AI, the group took a perspective at odds with the dystopian gloom that marks much discussion of AI impacts – and also, for that matter, at odds with the mixed outcomes that attended

by: Edward Parson, Alona Fyshe and Dan Lizotte

the actual technological and economic transformations of the industrial revolution. Instead, the group asked whether advances in AI could drive transformative social and political progress – and if so, what conditions would be necessary or helpful in promoting such progressive impacts. The group considered technical and socio-political conditions separately. Are there particular technical characteristics of deployed AI systems that would be most compatible with the aim to increase rather than decrease broad human liberty, equality, and agency? And what social, political, and economic conditions – including the need for viable business models – would be most conducive to AI systems with these beneficial characteristics being successfully developed, deployed, scaled, and sustained over time?

Regarding technical characteristics, the group identified two areas that might promise greater, and more broadly distributed, societal benefits than present and projected AI development patterns, one related to the structure of decision-making and one related to the scale, decision scope, and number of separate AI systems. Most methods of algorithmic decision-making, whether modern machine-learning or earlier approaches, structure their decision-making with the aim of optimizing a single-valued objective or scoring function under a single characterization, deterministic or probabilistic, of conditions in the world. An alternative approach, rooted in concepts of satisficing, bounded rationality, and multi-criteria decision making, instead pursues decisions that perform acceptably well under a wide range of possible realizations of uncertainties – and also under a wide range of plausible objectives and associated values. The group speculated that such robustness to diverse conditions is likely to be associated with greater pluralism of values, and with a tentative approach to decisions that recognizes uncertainty and limited knowledge, makes informed guesses, and seeks additional guidance – and thus, perhaps, with more inclusive and more equitable AI-driven decision-making.

by: Edward Parson, Alona Fyshe and Dan Lizotte

Regarding scale and scope, most present AI-based products are developed by for-profit enterprises and marketed to users – individual consumers, businesses or other organizations, government agencies, etc. – under conditions of asymmetric information and substantial market power. Moreover, users' values and preferences implicated by the AI systems are often under-specified, ambiguous, and manipulable, and may also exhibit systematic disparities between immediate impulses and considered longer-term values and welfare. The relationships between AI systems and users are thus ripe for exploitation to benefit the dominant party, e.g., by bundling attractive services with subtle, hard-to-observe costs such as loss of privacy or autonomy, or by manipulating users' adaptive and labile preferences to their detriment.

Many alternative models for AI deployment are plausible, at a wide range of scales in terms of people served and decision scope, and are potentially compatible with better advancing the pursuit of individual well-being and shared values. But achieving this alignment will require certain conditions, once again mainly related to the specification of objective functions but now with additional complexities that arise when multiple actors' interests and values are implicated. Such complexities include, for example, typical mixtures of shared, rival, and conflicting interests among actors, as well as collective-action problems and other pathologies of collective choice. In all such settings, AI systems must be faithful servants – which aim to advance as best they can the values and interests of the individual or collection they serve, even when these are tentative, imperfectly understood, and require continual adjustment – but with no consideration of the interests of the agent who developed or applied the AI.

Even if or when the associated technical requirements are clear, systems with these attributes may well not be compatible with present AI development business models.

Such systems will need contextual conditions that allow them to be developed, deployed, adopted, and scaled – while maintaining fidelity to the progressive aims and principles of the endeavor. The group worked through various scenarios of conditions that could enable such development, allowing the desired systems to gather initial development resources; secure the ongoing inputs needed to scale and progress; avoid being destroyed or corrupted by competition or attack from incumbents whose rents are threatened; and operate sustainably over time. Promising directions included a mix of strategic identification of initial targets; strategic early deployment of philanthropic or crowd-sources resources using open-source development; building strong early competitive positions through aggressive exploitation of IP advantages, coupled with binding pre-commitments to relinquish these at some certain future date; and compatible public policies regarding data ownership, IP, antitrust, and related matters. The group recognized that they were engaged in hopeful speculation about potential technical capabilities and associated societal conditions and impacts, when these conditions remain largely unexamined at present. They concluded, however, that in view of stakes and plausibility, these development directions merit high-priority investigation.

Concluding reflections: Routes to progress in understanding and governing AI impacts

As these short previews suggest, the discussions and outputs of the summer institute's work groups were too broad-ranging and diverse to admit any single summary or synthesis characterization. Still, a few salient themes emerged across multiple groups, including the following:

– The pluralism and ambiguity of values often embedded in current procedures, practices, and institutions, which may be put at risk by automation or codification of decision-making that exclusively optimizes for a single value – whether this single

value is efficiency or cost-minimization as often proposed, or something else;

– The rapidity with which considerations of AI deployment and impacts moves from seemingly prosaic considerations of system and application characteristics, to engage deep, even foundational questions of social values, political organization, and human identity;

– The frequency with which new configurations of responsibility and authority, in which AI-based systems augment and partner with human decision-makers rather than replacing them, appear superior on multiple dimensions to either human or machine decision-makers operating alone;

– The value, in considering ill-posed problems marked by deep uncertainty, of taking a dialectical approach – or alternatively, an adversarial or "red team-blue team" approach. This was clearest in the work of the two groups that struggled, from nearly opposite perspectives, with the thorny problems posed by the widely proliferating sets of AI ethical principles. The rich counterpoint between these two groups was unplanned good luck that emerged from the process of proposing and selecting workgroup projects. These groups have not yet had the opportunity to respond to each other directly: they were aware of each other's work from the brief plenary check-ins, but given the intensely compressed schedule of the summer institute there was little opportunity for substantive interaction between groups. Each group's work is limited and incomplete, in line with the aims of this rapid-output publication – as indeed are the outputs of all the workgroups. Yet they are also powerfully mutually enriching, offering complementary perspectives on the urgent question of how to inject ethical considerations into AI system development in practice, each hinting at potential correctives to the limitations of the other. They thus provide great heuristic value informing concrete early actions on a problem that defies resolution in any single step.

– The urgent imperative of finding footholds for progress in efforts to assess and

by: Edward Parson, Alona Fyshe and Dan Lizotte

govern mid-term developments and impacts. This is the place where immediate concerns and conflicts that suggest obvious – if unavoidably incomplete – responses shades into potentially transformative impacts. Yet this is also where early interventions hold the possibility of high-leverage benefits, even despite the relative scarcity of attention now being directed to these problems and the profound methodological challenges of developing disciplined and persuasive characterizations of risks and responses.

In addition to the substantive richness of the discussions and outputs, the Summer Institute also represented an experiment in process that greatly exceeded our expectations, which we believe offers significant insights into how to stimulate effective conversations and collective activities that deliver real progress on wicked problems like AI impacts, governance, and ethics. We noted above the conditions that make understanding or practical guidance on these issues so difficult to achieve, despite the flood of attention they are receiving – including deep uncertainty, rapid technical progress, and fragmented knowledge and expertise. Given that the familiar approach of waiting until impacts are determinate is insufficiently precautionary, what types of activity or process might promise useful insights for assessment or governance action? There is obviously no determinate checklist available, but a few conditions and criteria appear likely.

– It is necessary to mobilize multiple areas of relevant knowledge, expertise, and method – both across research and scholarly disciplines, and between academia and multiple domains of practice – because the problems' tentacles extend far broader than any single community of inquiry or practice;
– It is not sufficient to bring suitably broad collections of relevant expertise together; it is also necessary to facilitate sustained, intensive interaction, in which people dig hard into each other's concepts, methods, terminologies, and habits of

thought – to avoid the common failure mode of interdisciplinary activities, superficial agreement without actual advance of understanding;

– The problem of AI impacts and governance is both fast and slow: rapid pieces of technical progress and reactions to them add up to transformative changes over decades. There is news every week, yet the problem is not going away any time soon. There is thus a need to broaden debate and build expertise along generational lines as well as on other dimensions, to integrate instruction and professional development with parallel efforts to advance understanding. (This is one respect where the parallels between AI and climate change are instructive: both issues combine processes that operate on a wide range of time-scales, although in climate change there is much better knowledge of the long-term behavior of the relevant systems.)

– Knowledge in the field is diffuse, provisional, rapidly evolving. There is not an established and bounded body of knowledge sufficient to create an expert community. Plenty of expertise is relevant, but little that is on-point, certainly none that provides clear guidelines for progress.

These conditions suggest there is a need to encourage collaborative discussion and shared work along multiple parallel lines, which in turn suggests a decentralized approach to convening collaborative groups with the range of expertise needed to generate and pursue specific promising questions and ideas. The same conditions also suggest a need for communication vehicles to share questions, insights, arguments, and ideas, which is informed by relevant research and scholarship but proceeds faster, and more provisionally, than normal conventions of research and scholarship allow. Thus, even with the existence of such communication vehicles, there is also a need to develop a culture and practice of substantively rich but quick exchange of ideas, even provisional and incomplete. It does defy academic

convention, but on issues like this, rigor and completeness may be the enemy of progress.

These requirements suggest that the Summer Institute was, with small exceptions, a nearly ideal model to advance understanding and capacity – on AI and society issues, and on issues that exhibit similar characteristics. Indeed, the power of the model for similar issues was substantiated by the success of another summer institute convened two weeks later by one of the organizers here on a different issue, the governance of geoengineering. It appears to be a powerful model, subject to various conditions related to selection or participants, available time, etc., which clearly merits further development and application. We wish we could claim to have been prescient in designing this process, but there were large elements of luck in the outcomes of the Summer Institute. Still, the results – both those experienced by participants within the Summer Institute, and those marked by these first-round outputs – strike us as astonishing, given their origin as outputs of less than two full days of intensive focused work by newly formed groups. This was an exciting exercise to be a part of, and we are deeply grateful to our faculty and participants – for the intensity, intelligence, and good will of their shared explorations, and also for their participation in this experiment and the significant intellectual courage they have exhibited in allowing their work to be disseminated here in this provisional, incomplete form.

1. M Kranzberg, "Technology and History: Kranzberg's laws," *Technology and Culture* 27:3, 544-560, July 1986
2. For vivid illustration of this point, see the recent collection of highly cogent, yet often mutually contradictory or incommensurable speculations on AI's trajectory and significance, in John Brockman (ed.), *Possible Minds: 25 ways of looking at AI*, Penguin: New York, 2019.

3. For a more detailed characterization and examples of this intermediate range of potential impacts, see Parson et al, "Artificial Intelligence in Strategic Context," at https://aipulse.org. See also Seth Baum, "Reconciliation between factions focused on near-term and long-term artificial intelligence," AI and Society 33(4):565-572 (2018).
4. The authors thank Summer Institute faculty members Trooper Sanders and Michael Karlin for discussions that identified and clarified these issues.
5. For younger readers, this group's title refers to Sir Humphrey Appleby, a fictitious senior advisor ("permanent secretary") in the UK government in the classic BBC television series, "Yes, Minister" and "Yes, Prime Minister" between 1980 and 1988. Both series provide deeply insightful (and hilarious) views of the relationship between advisors and political leaders in high-level policy decisions – and not just in Britain.
6. For an insightful recent discussion of these issues, see J. Whittlestone et al, "Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research," London: Nuffield Foundation, 2019, at http://lcfi.ac.uk/resources/ethical-and-societal-implications-algorithms-data-/.