by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

AI's Transformative Social Impacts and their Determinants

The potential societal impacts of artificial intelligence (AI) and related technologies are so vast, they are often likened to those of past transformative technological changes such as the industrial or agricultural revolutions. They are also deeply uncertain, presenting a wide range of possibilities for good or ill – as indeed the diverse technologies lumped under the term AI are themselves diffuse, labile, and uncertain. Speculation about AI's broad social impacts ranges from full-on utopia to dystopia, both in fictional and non-fiction accounts. Narrowing the field of view from aggregate impacts to particular impacts and their mechanisms, there is substantial (but far from total) agreement on some – e.g., profound disruption of labor markets, with the prospect of unemployment that is novel in scale and breadth – but great uncertainty on others, even as to sign. Will AI concentrate or distribute economic and political power – and if concentrate, then in whom? Will it make human lives and societies more diverse or more uniform? Expand or contract individual liberty? Enrich or degrade human capabilities? On all these points, the range of present speculation is vast.

What outcomes actually come about will depend partly on characteristics of the technologies, partly on the social, economic, and political context – what specific technical capabilities, with what attributes, are developed and deployed, and how people adjust behavior around the capabilities. It is a basic doctrine of technology studies to reject technological determinism: technological and socio-political factors interact, and to the extent either predominates in shaping outcomes it tends to be the social and political factors. The interplay between these underpins the well-

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

known "Collingridge paradox," which states a structural challenge to managing technology's societal impacts[1]: early in development, control efforts are hindered by limited knowledge, because impacts are indeterminate until a technology is stabilized, deployed, and used; while later in development, control efforts are hindered by limited power, because the same development processes that determine and clarify impacts also build political interests in the technology's unhindered expansion.

In correctly rejecting naïve or extreme forms of technological determinism, however, these characterizations are often deployed too starkly and universally. Collingridge's paradox of knowledge and control is better understood as a persistent tension than as a categorical statement of impossibility. Moreover, without disparaging the power of social context, technological processes and artifacts are not infinitely malleable: particular technologies have characteristics, which in some cases tend to favor particular uses, applications, or consequences. Kranzberg's (slightly whimsical) first law of technology aptly captures the tension: "Technology is neither good nor bad; nor is it neutral."[2] It is subject to influence, and that puts responsibility onto humans to wisely guide its development and application.

AI may be a class of technologies for which serious consideration of the role of technical characteristics in shaping impacts is especially needed, in view of its labile nature and its potential for profound societal disruption. Two examples from widely separated parts of present debates about AI impacts illustrate the point. First, concerns about impacts of extreme AI advances to general, beyond-human intelligence – and related efforts to develop "Friendly" or "Safe" AI, or align its objectives with human values (assuming these are known and agreed) – are entirely concerned with attributes of the technology. These efforts seek to ensure good

consequences, or at least avoid the worst ones, by embedding reliable determinants of benevolent aims, prudence, or other virtues into the technical artifacts themselves. To the extent this program succeeds – a huge assumption, to be sure – it would move concerns about these extreme forms of AI impact out of the social and political domains entirely.

It is widely noted, of course, that focusing predominantly on such hypothetical future super-AI risks misleading, by distracting from addressing nearer-term uses and impacts that are also potentially transformative for good or ill – including both the "now" and the mid-term.[3] Technical characteristics, even abstracted from social context, also matter for these near and medium time horizons – i.e., well before development of AGI or super-AI – when AI will clearly have transformative possibilities but still, at least formally, be under human control. The importance of technological characteristics is evident even in current AI controversies, in both what technical capacities allow and what they require. As an example of impacts driven by what technical capacities allow, AI-enabled advances in data integration and surveillance, especially facial recognition, already present significant threats to privacy and autonomy. These capabilities are being deployed because current actors find advantage in them, of course – a matter of social and political context. But it is the technical performance characteristics that create these new capabilities and make them visible. As an example of impacts driven by technical requirements, present machine learning algorithms require training on large labeled datasets. This requirement has driven two powerful effects and points of concern. It has steered many near-term commercial applications toward decision domains such as criminal justice and health, in which huge individual-level datasets with clearly labeled outcomes are available, with little advance consideration of the high personal, legal, and societal stakes – and high costs of error – that are intrinsic to these domains.

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

And it has replicated, by some accounts even magnified, pre-existing biases present in these training data, and projected them forward into future decisions.

Our workgroup reflected on the question of AI impacts in broad historical context: in effect, we took seriously the analogy to past technology-fueled revolutionary transformations of human society such as the industrial revolution. But we did this with a perspective opposite to much current debate, considering the prospect for societal impacts that are transformative in scale but beneficial in valence. Speculations about huge societal benefits from AI are common, but tend to be superficial and conclusory, often based on speculative gains in single areas such as medical care or scientific research. By contrast, speculations on AI-driven dystopias are frequent and attention-getting, often with their causal mechanisms characterized in some detail. [4]

A Historical Analogy

In our inquiry, we drew insight and inspiration from a line of commentary on past societal transformations that gets insufficient attention in current debates on technology impacts – despite being a prominent theme in the work of a few distinguished scholars such as Albert Hirschman and Elizabeth Anderson.[5] These scholars point out that at the time modern liberal states, market capitalism, and associated technological changes were emerging, these trends were widely heralded as drivers of political and economic progress, relative to aristocratic social hierarchies, promising not just greater liberty – one part of the argument that remains prominent in modern political discourse – but also increased equality (in some accounts also fraternity or comity – to complete the revolutionary triad). These promised and briefly realized happy trends reversed, as technologies of the

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

industrial revolution and their economies of scale drove vast accumulations of capital and separated the previously tight connection between markets and equality. Progressive reactions from governments (e.g., anti-trust) and new organizations (e.g., labor unions, charitable foundations) mitigated these trends to better balance autonomy, prosperity, and equality – a balance that current technological and economic trends are disrupting.

Our group aimed to re-open this question in the current context of rapid advances of AI. Can these transformative capabilities deliver on the old promise of technology as both liberator and equalizer? Can they do so in a way that is compatible with foundational moral and constitutional principles, and democratic institutions: e.g., freedoms of speech, association, religion, and the press; and private property rights with markets allocating resources through voluntary transactions, except insofar as these implicate external harms or public values (and as Mill reminds us, without drawing these public bases for concern so broadly as to undermine the basic liberty presumptions).[6] And if this is all possible, what would it require: what are the key conditions that would mediate the ability of AI to help advance such a happy social vision?[7]

In considering this question, we did not elevate technological characteristics to the exclusion of social and political context; but we did consider technical and political forms of the question separately. First, what technological characteristics of AI systems and applications are likely to promote good societal outcomes? And second, what economic, social, and political conditions – including, concretely, what feasible business models – are likely to promote AI technology developing in these benevolent directions, and be sustainable over time?

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

Promising Directions: Technological Characteristics

In considering the technological part of the question, we focused on two broad technical attributes that we speculate may help direct AI's transformative societal impact toward the good: one related to the form and structure of decision-making, and one related to the distribution, scope, and number of separate AI agents.

Decision-making structure: Single-valued optimization, versus robustness and pluralism?

Virtually all automated decision systems – modern machine-learning systems and conventional algorithms alike – operate by optimizing a single-valued scoring or objective function. This is most obvious in the case of known preferences and conditions of full certainty, but similar approaches are used under uncertainty: maximizing an expected payoff or expected utility function, based on specified probability distributions, sampling from specified uncertain parameter inputs, or data assimilation from concurrent observations. These approaches all optimize a single-valued function relative to a single characterization, deterministic or stochastic, of conditions in the world.

There is an alternative, less unitary approach to decision-making, which initially grew out of concepts of satisficing, bounded rationality, and multi-criteria decision-making.[8] This alternative approach, one prominent form of which is called "robust and adaptive decision-making" (RDM), seeks decisions that perform acceptably well over a wide range of possible realizations of uncertainties, rather than performing maximally well under any single specification, whether deterministic or probabilistic. RDM has extensive experience in diverse decision application areas. It has not been

used in AI or machine learning, but we conjecture that it may have powerful implications, broadly consistent with the progressive social values we aim to advance.

The seed for this hopeful speculation lies in the fact that RDM is not just robust over alternative realizations of uncertainty about the world: it is also robust to uncertainty in the decision's goals or the range of values it implicates. RDM thus holds the potential to be more pluralistic, more compatible with both uncertainty and diversity of values – and thus, perhaps, with more inclusive and more equitable AI-driven decision-making. We realize that this is hopeful speculation about potential technical capabilities and the societal implications of their application, not a demonstrated characteristic of AI systems. But while the capabilities and associated questions remain largely unexamined, they clearly merit high-priority investigation.

The Number and Orientation of AI agents: What actors, and what aims, do they serve?

There is a wide range of speculation on the number, deployment scale, and objectives of future AI systems, ranging from each person commanding multiple AI agents for different purposes, through one integrated AI that does everything for everyone. Present AI developments consistently show a much narrower pattern, which is not necessarily well aligned with broadly distributed societal benefits. Most current efforts and most important recent advances have come from large, well-funded organizations: for-profit corporations, free-standing laboratories and institutes, and universities, some surrounded by clusters of small startup firms, with widely varying levels of government financial support and control among countries.

The most prominent current deployments of algorithmic decision-making are offered by private, for-profit firms, many of them in settings where the deploying party has a

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

dominant position in the relevant interactions: Amazon toward purchasers and 3rd-party vendors; Facebook toward social-media users; Google toward its service users and data providers; Uber toward drivers. In these settings, users can observe only a small slice of the system's performance in its interactions with them, but have virtually no information about its broader operations, including what it is optimizing. In all these interactions, commercial or not, the available evidence – plus common sense – suggest that the systems are optimizing for the interests of the dominant actor, taking lesser account of the interests or welfare of the user only as needed to advance the primary aim – and, moreover, are doing so in ways that take advantage of the dominant actor's market power.

But this structure of relationships is not a necessary consequence of algorithmic decision-making or decision-support systems. One can imagine a wide range of other possibilities for how AI systems are deployed, some of them more compatible with a reduced concentration of power. An obvious and widely discussed possibility would be general-purpose AI assistants serving individual people, either unitary systems or integrations of multiple special-purpose systems. Such agents could act as information source, advocate, and negotiator for their clients in multiple interactions. They could provide suggestions and recommendations, manage the mechanics of transactions, and bargain on your behalf in consumption and other commercial interactions, both present ones and new ones it would enable – e.g., renting out your car, tools, or other costly assets when you do not need them. They could play similar roles in financial and investment decisions, and in labor-market participation. In situations of conflict or interaction with authorities, they could aid you in negotiations and advise you on your legal rights. They could support and advise your political participation, whether through existing channels such as voting and candidate support or through new, AI-enabled processes that combine elements

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

of representative and direct democracy, such as issue-specific proxy delegation or other forms of "liquid democracy." And they could act as a personal coach, helping you make decisions and manage your time in line with your goals and values. The biggest challenge in creating such systems – as we discuss below – would be defining their objectives to reliably align with their user's welfare and values.

Alternatively, rather than serving individuals, AI systems could operate enterprises or collections of assets, to perform specified functions or advance specified interests aligned with social good. For example, AI systems – perhaps self-owned or self-controlled – might operate businesses or parts thereof such as individual factories; apartment buildings or larger-scale collections of housing or other buildings; public transit systems or other infrastructure components; or specific functions of government decision-making, in cases where the delegation of authority and the specification of relevant values to advance are unproblematic.

AI agents could also be deployed at higher levels of aggregation, to inform or guide the joint actions of groups of people in pursuit of their shared interests in some specific domain – whether commercial, political, recreational, expressive, religious, or something else. Relative to personal AI assistants, these agents would operate at a scale that is broader in the people whose interests are served, but narrower in the range of functions being pursued or interests being advanced.

Finally, one could imagine AI agents deployed at the level of the entire polity in some jurisdiction, centralizing decision-making on state functions in pursuit of some legitimate and widely agreed conception of the aggregate social good. There is of course some tension in using AI this way to increase human liberty and agency. Can we really claim to advance liberty and agency by centralizing state control? But these are largely the same tensions as attend state authority guided by humans. The

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

state is a strong centralizer of power. But in liberal democratic states, this centralization serves the interests of order and security, including displacing other, less legitimate forms of concentrated power that are likely to arise in the absence of the state. And moreover, liberal states exercise this power lightly, so as to enhance liberty and welfare, only coercing citizens as needed to pursue legitimate public purposes.

Considered overall, this collection of potential AI deployments might tend to have an hourglass structure. As the scale of deployment moves from individuals to groups, the functional scope of the AI narrows to specific aims of particular groups; then at the highest level of jurisdictional aggregation, the scope of AI decision-making returns (or can return) to the comprehensively broad, imperfectly known set of interests that are the legitimate purview of state authority.

Examples of large-scale social reorganizations that would potentially be feasible with such a collection of AI agents would include the following:

- Breaking up monopolistic social-media platforms into multiple distinct platforms, each managing members' interactions by internally agreed rules and mediating the interactions between insiders and outsiders. AI would be used to facilitate such a breakup by overcoming the incumbent advantage due to network externalities. This is already happening in a small way, with the growth of new social media platforms such as Diaspora and Mastodon with commitments to stronger privacy protections than present platforms.

- New ride-sharing platforms, in which AI is deployed not as an instrument of the network's monopoly (or the Uber/Lyft duopoly) over drivers, but instead deployed to serve drivers and driver collectives, interacting with multiple counter-parties (current ride-share companies, potential new entrants, and others), and

with riders, who in turn might be interacting with the system through their personal AI assistants. AI in such settings could optimize contractual terms to maximize shared value, while also equitably distributing surplus value between drivers, riders, and providers of other factors of production – including paying returns to capital and managerial services, but on competitive rather than monopolistic terms.

– A similar but broader labor-force model, not for people providing one service to one business (e.g., drivers and Uber), but with AI intermediaries enabling groups of individuals to come together to offer products and services to the market without the need for corporate intermediaries. Such groups could utilize distributed supply chains, to which they submit opportunities and find others who wish to join and bid to offer their services. As in the narrower, ride-sharing case, AI agents could optimize contractual terms, including provisions for duration and modification, based on the preferences of the participating individuals.

– AI-mediated political interaction – among citizens, activists, politicians, and political parties – to provide more civil and substantive deliberations, and more effective, informed, and flexible translation of citizen preferences into collective decisions. Such systems might aim to provide equal opportunity for political participation to all citizens, to motivate and reward virtue and moderation rather than vice and extremism, and to be dynamic – able to update, including updating collective understanding of what counts as virtue or vice. In contrast to the preceding examples, which would replace commercial transactions, this one would require more manipulation of incentives for individual behavior in pursuit of collective interests of civility, moderation, and reasoned debate. It could be designed to reward – with more influence and scope to reach a broad audience – those who best exhibit those virtues, rather than rewarding volume, belligerence,

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

> extreme views, personal attacks, skilled manipulation, or outrage.
>
> – Within this context of AI-mediated political interactions, AI could discharge certain administrative functions of the state, mitigating the long-standing tension between expert and democratic control noted by Weber. AI's exercise of these functions would be guided by objective functions tuned by democratic deliberation. Through AI-facilitated direct deliberations or some equivalent quasi-legislative process, citizens would define large-scale aims and principles, set parameters for AI objective functions, then observe the results and iteratively adjust those parameters to steer toward a preferred balance of multiple societal aims. Such an administrative AI would in effect act like regulatory agencies under present administrative law, but with more explicit and more consistent parsing of authority between high-level democratic goal-setting and technically skilled implementation.

Realizing any of these alternative models of AI deployment would pose major challenges, which include significant technological elements even though they are not exclusively technological in character. A central challenge, perhaps the fundamental one, is appropriately defining the AI's objective function. Even with the shift toward a more robust and pluralistic approach as discussed above, this would imply three additional subtle and related requirements.

First, in any application the AI must act as a faithful agent of its intended beneficiary, whether this is an individual or a group of any size. The AI pursues its beneficiaries' values and interests – not the interests of its maker, not even when it must resolve ambiguities or indeterminacies in its understanding of its beneficiaries' values. This would represent a major departure from presently deployed AI systems, including those that are approaching the role of general-purpose personal assistants.

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

These are developed by firms with interests in the user's behavior, and thus in manipulating that behavior or harvesting the user's information – even if that manipulation may be subtle and the systems seem to optimize for the user's preferences. These systems are also developed in the context of various commercial and state interests in creating over-rides or back doors, in order to allow surveillance and control contrary to the user's interests.

Even assuming this first condition is met, so the decision scales are not tilted to favor the maker's interests, systems interacting with individuals face a second challenge of understanding the determinants of the user's true values, interests, or welfare, as distinct from their immediate impulses or desires. This is hard to define, imperfectly inferable from observed behavior, prone to error, and in need of continual adjustment and updating. Like a wise parent or a skilled life coach, such systems would nudge the user's choices in directions judged likely to be compatible with their long-term flourishing – with the key difference from parenting (although not from coaching) that the ultimate authority in the relationship lies with the user. This would require a delicate balance, by which the system pushes against immediate preferences and desires when these appear to be at odds with the client's values or long-term interests. But to do this, the AI assistant must build a model of the client's values and long-term interests, based on data available to it. The system will thus sometimes make mistakes, and so will need to recognize uncertainty, make some of its recommendations tentative, and sometimes consult and ask for help – while also still using its present, uncertain knowledge to configure the choice space in ways likely to tend to beneficial outcomes. There will thus be a core design tension, between allowing human over-ride of AI recommendations and putting some degree of burden or barrier in front of instant, effortless, or wholesale over-ride.

A related but even sharper tension will be present in the case of people with

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

destructive, malicious, or criminal preferences. Even liberal states do not honor or aim to fulfill the preferences of every citizen, independent of collectively exercised moral judgments. One can readily imagine a sexual predator or other criminal wanting their AI agent to help identify victims, assess the threat of detection or apprehension, manipulate victims to not resist or not report, or pursue other clearly nefarious aims. One problem here is defining the boundaries of permissible preferences – a challenge similar but not identical to that in non-AI contexts of defining the boundaries of criminal or civil wrongdoing – except that, as in so many domains, making scoring or objective functions explicit can be troublesome in cases where maintaining ambiguity provides needed social cohesion or moral comfort. Even assuming appropriate definition of the boundaries of permissible user preferences, a related design problem will be protecting AI systems against hacking or manipulation to enable such uses – either by intentionally disabling the AI's "conscience" functions, or by misrepresenting intentions in planning or multi-step execution of bad acts. We want individual AI agents that can distinguish their user's seeking an out-of-the-way place for a quiet picnic, or to carry out a murder.

Additional requirements and challenges would apply to AI agents managing enterprises or assets: e.g., self-directed AI corporations, housing developments, or transit systems. First, should the substantive decision scope of such agents be narrowly circumscribed and fixed? This raises issues analogous to those in current law regarding charities or other non-profit organizations seeking to change their original missions. Narrow and fixed goals would risk restricting behavior so the AI cannot respond appropriately to changed knowledge and conditions; but changeable goals risks letting the AI transit system decide to go into the adult film business instead – whether because it judges the change would make more money, generate more happiness, or better promote peace and order. Second, can behavior be

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

constrained to be legal and ethical in a way that is sufficiently clearly defined and does not put such enterprises at competitive disadvantage relative to others playing by looser rules? Third, can objectives be tuned to not accumulate rents in excess of the costs of all factors of production? If so, these enterprises might be able to out-compete others that are pursuing and taking rents, and so form the kernel of a gradual erosion of concentrated economic power – unless the others are pursuing an Amazon strategy, taking losses for a long time to secure a dominant market position thereafter. Alternatively, if rents do accrue – as they sometimes will – what should be done with these? Presumably they should not be retained within the individual enterprise, but instead distributed in line with the system's large-scale aims. But does this mean to the Treasury? Or perhaps to a pool dedicated to financing the capital needs of the broad "social-progress-through-AI" enterprise, as discussed below? Finally, if these bodies sometimes go bankrupt – as seems likely, given the constraints imposed on them – how can one ensure that they quietly accept this fate, and what should happen to their assets when they do? As UCLA law professor Dan Bussell argues in a forthcoming paper, AI enterprises may need a new kind of bankruptcy court.

When AI systems are deployed to serve multiple people, to inform people's interactions with each other or advance group interests and values, additional challenges and design tensions will arise. These problems are similar, whether the structural approach to decisions involves collective decision-making or bargaining among individuals' AI agents, or some separate AI agents operating at a higher, collective scope of authority. The challenges all follow from a basic fact: in any decision situation involving multiple people, there are multiple measures of welfare. These are sometimes aligned, but they can also exhibit disagreements, rivalrous claims on the same resources, collective-action problems, or other tensions. Most

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

often, there is some mixture of aligned and opposing interests. In such situations, even formal game-theoretic outcomes can be ambiguous due to the existence of multiple Nash equilibria. There can also be inferior collective outcomes from individual choices that are locally advantageous, or inequitable outcomes in distributive negotiations that favor the most aggressive bargaining tactics.

Even assuming AI agents reflect individual values well, guiding or informing such multi-person interactions presents several additional design requirements. The systems would need to identify and avoid collectively inferior outcomes – even if they are equilibria – by providing coordinated nudges to steer parties toward collectively superior outcomes. They would need to apply the same gentle resistance against self-destructive impulses as at the individual level, now with the added requirement to steer groups against choices driven by collective-level pathologies such as envy, malice, hostile stereotypes, or escalation dynamics and other entrapment mechanisms. And they would need to address the problem of aggressive bargaining behavior, recognizing that this often succeeds at securing favorable one-time outcomes in divide-the-pie negotiations, at the cost of inferior collective outcomes and damaged relationships. The systems would have to both refrain from such behavior on behalf of individuals, and not reward it in determining collective outcomes. These requirements and the associated bargaining pathologies are best understood in commercial interactions, but they have close analogies in other domains. A salient current example is maintaining civil discourse, in politics and online, in the presence of powerful attention-getting advantages in being colorful, extreme, and uncivil – a domain in which a few experiments have shown that AI agents can make the problem worse, if they are trained on the actual content of current discourse.

Achieving these aims would require that an AI system managing collective decision

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

outcomes would need both the knowledge to identify collectively superior and inferior outcomes, and the ability to apply defensible principles for fair division of surpluses and resolution of conflicting preferences. If collective decisions are handled by collectively accountable AI agents, these would need to reliably observe the preferences and values of all affected people, plus relevant information about the world that shapes the set of feasible outcomes – a tall order. On the other hand, if collective decisions are handled by interactions among individual AI agents – each presumably with better information about its own user's preferences and values — then the individual agents' bargaining behavior must be subject to constraints guided by collective welfare: e.g., seeking to maximize joint gains; not pursuing these by shifting negative externalities onto others not present in the interaction; fair dealing with each other, in both process and substance; and refraining from destructive bargaining tactics even when these promise a one-time advantage.

Some form of regulation at the collective level appears to be needed, but defining (and automating) precise rules will pose severe challenges. In different decision domains, the needed functions might be characterized as mediator-arbitrators, content moderators, or judges. Should these be AIs, humans, or machine-human partnerships? How can these processes be made robust against sophisticated attempts to capture them for partisan advantage? If the aim of these is to advance widely (but perhaps not universally) held collective values, how broadly should they be binding in domains such as political discourse that implicate free speech and other liberty values? And to the extent these processes supplant human decision-making – which traditionally advances collective aims by some combination of formal regulation and propagation and maintenance of social norms – might widespread assumption of these duties by AI risk atrophy of the associated skills, sense of duty, and other virtues in humans?

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

Promising Directions: Social, Political, and Strategic Issues

Summarizing the above, the technical AI characteristics we speculate likely to be associated with good societal impacts include the following:

- AI does not irreversibly alienate individual human agency in any domain;

- AI objective functions are tentative and pluralistic, along the lines of RDM, rather than single-minded and dogmatic; they admit multiple possibilities in outcomes and values, recognize limits to their knowledge of these, and know when and how to ask for additional information or guidance;

- AI performance is monitored and adjusted over time with significant input from people, acting alone for their personal AI's or in democratic, deliberative groups for AIs with collective or society-wide responsibilities;

- AI agents must be are trustworthy in all respects. Individual AI agents pursue the interests of their client rather than any developer or vendor; and they pursue the true, long-term interests and values of their client, via recommendations, nudges, and exhortations – acting like a wise parent or friend. AI agents acting, mediating, or arbitrating on behalf of collections of people follow principles of fair dealing and equitable distribution of surpluses among participating parties, and incorporate interests of other actors or values outside the participating parties only insofar as these represent real externalities.

Having speculatively identified these desirable technical characteristics of AI systems, we then asked how such technical systems might be developed, deployed,

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

scaled, and sustained over time. These are questions of political and economic strategy. The proposed innovations – in addition to being uncertain and weakly characterized – would represent attacks on existing concentrations of wealth and power and the rents that sustain these. There are thus likely to face, at a minimum, challenges in securing the resources they need to be created, established, grow, and sustain; and more likely, will face determined and strategically sophisticated opposition.

Getting a Start:

In this situation, the first challenge will be getting such systems developed and deployed. What this requires will depend on the details of the relevant systems and the inputs needed to produce them – the production function for AI capabilities – all aspects of which are deeply uncertain.

On this, an initial issue to consider is whether systems with the desired characteristics can be reliably developed by modifying other systems that were developed by and for current commercial actors – assuming these can be legally acquired. If they can to some degree, then the key questions are, first, trust and reliability – how can we verifiably assure that the systems so ported do not sneakily import the interests of their developers – and second, what additional resources and inputs are needed to modify systems and deploy them for their new purposes?

At best, the desired systems would need training procedures and data for their newly targeted uses, related to the individual or collective values to be served. This might be cheap and easy; it might be expensive and difficult; or it might be impossible, at least initially, because data relevant to the newly targeted uses and goals might not exist. Oddly, there is likely to be more and better data available to serve vendors' commercial interests – which depend on observable matters such as

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

attention, time spent, and purchasing and other behavior – than is available to serve individual and collective values. Data presents other challenges as well, including the possibility that no truly general-application AI can be developed given jurisdictional divisions and restrictions on data access and use; and the present dependence of AI progress on a huge volume of labeled data, which in turn depends on a huge, low-wage workforce doing this essential step.

The less fortunate case would be that new systems with the desired characteristics must be developed from scratch. In this case, the same data concerns identified above would still apply. But there would also be a greater need for other inputs, for initial system development and deployment and for continuing maintenance, adaptation, and upgrades. These needs are probably similar for key advances in multiple areas of AI development, independent of the specific form of objective or the scope of application. In addition to suitable training data, these include highly skilled technical personnel; hardware-based computing power; and capital – lots of capital judging from present industry structure, although this could change.

The premise of the new AI developments we seek is that, unlike the present system, successful development of useful capabilities, even achieving crucial technical advances, will not create fabulous wealth for developers or their employees, collaborators, or investors. So how can the needed developments be effectively motivated? The recent case of OpenAI reconstituting itself as a for-profit corporation because it could not raise enough capital as a not-for-profit AI developer provides a germane cautionary example.

Our discussions identified several promising elements of potential development models. The first concerns identifying early targets, current products or present or potential uses to displace. Promising targets might include products that are now

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

gathering the largest rents, or that are targets of the strongest current objections and political threats, or for some other reason are ripe for raiding. Other promising factors would include consumers' willingness to incur a little inconvenience from switching costs; perhaps also a preference for local providers and small-scale relationships. The aim would be to target early penetration there, with alternative products that distribute the rents or other values to their users, not the vendors.

The second element is assembling and mobilizing the needed factors of production. On this, the initiative could start with crowd-sourcing, philanthropy, or other sources of capital motivated by social goal rather than profit – although these sources are usually much smaller than investment-motivated capital. An open-source development model may hold advantages, including facilitating engagement of top technical talent and mobilizing utopian and anarchic strains within the technical community. Such an initiative would provide an opportunity to probe the depth and sincerity of the revolts by high-tech workers against narrow conceptions of their employers' self-interest, inviting them to put their money and skills where their mouths are.

All aspects of this strategy – including, crucially, attracting capital and *pro bono* talent – would benefit from well-branded, highly attractive initial projects: e.g., the faithful individual AI helper, or the AI facilitator of civil political discussion and collective action (both of which may represent compellingly attractive aspirations, but would clearly need better names).

Not all philanthropy pursues aims that are clearly benign and universally agreed, of course. Sometimes it makes sense to worry about limited or partisan social objectives in philanthropy: For example, don't solicit support for your climate-change campaign from the Koch Foundation. But this concern might be less serious

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

for AI than in more established policy areas with well-known lines of political alliance and opposition. Libertarian philanthropists – yes, perhaps even the Koch Foundation – may well support the aim of empowering individual liberty and agency with individual-level AI agents. As for group-level AI agents advising different decisions to advance different aims, these will be multiple overlapping agents operating in a pluralistic setting, so the risks of capture by any limited or partisan view of the public interest may be less severe.

Persisting and Scaling:

Once socially beneficial AI capabilities are deployed, they still need mechanisms and resources to persist, scale, and sustain their position. Moreover, they must do this in a way that maintains their alignment with citizen and public values and remains attractive to users – even once the initial novelty of the initiative has passed, with possible decline in the enthusiasm of pioneer supporters and developers. The initial sources of capacity may not be enough to persist under these conditions, or to overcome the sustained advantage of strategically sophisticated and ruthlessly self-interested incumbents, who might respond by deploying cheap attractive systems as loss-leaders to secure longer-term advantage. The enterprise will need to maintain needed access to technical expertise and capital, whether from associated revenues or from investors.

Some present business models, such as relying on advertising, clearly appear not to be viable for this project, but several others appear plausible. One possibility would be subscription or purchase, although the implications of alternative ownership models and their compatibility with the large-scale aims – do I purchase my AI assistant and related supporting systems or rent them, and from whom – would require careful thought. If the services provided by AI systems include facilitating

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

transactions or cooperative activity with exchange of money, the system could take a fee to cover development costs, provided the fee is perceived as reasonable and its basis fully disclosed. Another possibility would be a co-operative enterprise model. These organizations reach large scale in some jurisdictions with strong historical traditions of self-organized cooperative activity and supportive policy environments.

There might be bootstrapping possibilities, based upon the use of AI. Early AI's might be developed to help identify targets and strategies for subsequent expansion. They might provide information, services, and access to resources that have traditionally been provided by venture capitalists or other early-stage private investors. They might help identify points in current supply chains or production models that are rigid or constrained, or where market power is hindering rapid development and deployment.

Another novel approach might be to turn the widely denounced short-termism of capital markets to advantage, by deploying AI agents that pre-commit to change their behavior over time. An AI raider could initially pursue maximum short-term competitive advantage, but with a binding commitment to change course in the future. If its short-term competitive advantage is based on strong IP, for example, the commitment might be to unlimited free licensing after the initial period expires. A policy change to support this might be a new form of IP, based on modifying either patent or copyright, that combines highly advantageous short-term protections with an iron-clad, non-contestable commitment to expiry and full release to the public domain thereafter.

As the endeavor succeeds and grows, it will encounter changes in its strategic and competitive conditions. Some of these will work to its benefit: for example, open networked organizations pursuing broadly public aims are likely to have an easier

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

time pooling and sharing data than rivalrous commercial organizations. Other changes will represent new challenges that increase costs or other barriers. As the new systems grow to mediate decisions that channel large sums of money, they will attract hackers and others interested in subverting them, and will have to develop robust security protections. Stringent open-source review can provide part of the needed protection, but some risk will remain. It will also be necessary to be vigilant about the interests of continuing sources of finance: any source motivated by financial return will present ongoing risks of subtle distortions of aims, and the associated prospect of simply replacing old centers of concentrated power by new ones just as determined to sustain their position.

Finally, if the endeavor succeeds so well that some combination of individual AI assistants, autonomous AI enterprises, and AI-mediated collective interactions – all with the desired characteristics – becomes the dominant model for societal deployment of AI, it will be necessary to grapple with the question of innovation. Current law and policy assume that the main incentive to innovate comes from the pecuniary motive of earning rents, from the innovations themselves and from IP protection around them. With AI agents eschewing most or all of the rents that provide enormous financial rewards to present market actors, where will the motivation and resources to support innovation come from? Several alternatives might be possible. Innovation might still come from people, businesses, or other organizations, including AI-facilitated innovation, stimulated by some combination of the pecuniary rewards that remain under the new model (which will be smaller than under the present system, but probably not negligible), plus intrinsic motivation to innovate and create – which the present system largely overlooks. AI agents might be able to fully take over the huge volume of prosaic, small-scale innovations now done for profit in enterprises seeking IP assets – many of small or questionable merit. AI

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

agents could take over the pedestrian activities of searching through current technologies, patents, and scientific publications that power much such innovative activity, but do so with better information and processing capability and with objectives better aligned with the broad public interest – and with results placed in the public domain for free further exploitation. For larger-scale scientific, technological, artistic, and social innovation, intrinsic motivations have long been the dominant driver and it is reasonable to expect they will still be present in the new world. Indeed, they might be effectively aided by AI support tools.

Challenges, next steps:

The technological-political program of societal transformation we sketch here is bold, under-specified, and incomplete. It can be viewed as an attempt to update Alinsky's Rules for Radicals, Scott's Weapons of the Weak, and the Ethical hacking movement, for a new technological environment of greatly increased power for autonomous and semi-autonomous systems. It is bold in that we are proposing a new technological model of AI and its deployment that opposes the interests of present dominant incumbents – both private-sector actors whose revenues and business models would be threatened, and government institutions that would hold different and less extensive and exclusive as some decision authority shifts to networks of citizens and autonomous decision-making systems. It is essential not to be naïve about how large the barriers to entry are, or about the determination and resources of incumbents seeking to strangle the new model in its crib. The new model also opposes certain structural characteristics in the economy that tend to favor scale, and thus centralization. These include technical factors such as economies of scale and network externalities that are strongly shaped by characteristics of production technologies; and factors more institutional and political in origin, such as fixed costs

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

from regulatory obligations advancing various public values such as environment health and safety, consumer protection, etc.

The new model is under-specified, both in its technical and its political/strategic dimensions. Technically, we sketch a couple of salient system characteristics that appear likely to push in the desired direction, but the devil is in the details. A wide range of systems and design approaches is now being pursued and developed in parallel, with capabilities – depending on multiple factors in the systems and their contexts – that might favor or oppose liberty, privacy, agency, and equality. Even current developments have had a mix of centralizing and decentralizing effects, empowering many distributed activities even as they create great new centers of wealth and power, including new forms of power not yet exploited or even well understood.

As a strategy to move toward this vision, we have identified a few possible pathways to pursue it through private action. But it is also worth asking whether appropriate government policies would be necessary or helpful, and if so, what form. Possible points of leverage might include data ownership policies such as clear conferral of data property rights on individuals, or limits on concentrated holding of data; limits on or new forms of IP; or more expansive definition and robust enforcement of anti-trust policies. To the extent the desired transformation does require public policy, one might also consider which jurisdictions would be most promising to seek an early strategic foothold. Perhaps the social democracies of Europe, which are already leaders in data and privacy policies? Or perhaps major developing countries with strong technological capacity – who would have the advantage of large domestic markets for early scaling, but might also be ambivalent toward the leveling ambition, depending whether leveling is construed as between countries (in which case they would presumably be keen advocates) or within countries (in which case, maybe not).

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

In this global context, one must also consider risks posed by opportunistic geopolitical adversaries, including the possibility of surreptitious early support for the development of the new systems coupled with efforts to bias or undermine its aims – although this threat might become less salient over time if one consequence of the spread of the new systems is a decrease in international rivalries.

Finally, the proposed new model is incomplete. It is unlikely to address all impacts and social disruptions caused by rapid advances in AI. In particular, we can't necessarily expect it to avoid large-scale displacement of livelihoods by AI. It might, however, make mass unemployment less individually and socially destructive, perhaps even make it desirable. If leveling of power implies different bases for distribution of economic output, no longer coupled to employment, then loss of employment might cease to be catastrophic. This might seem inconceivable, but it could be analogous to the treatment of health insurance across nations: in the United States it is tightly coupled to employment and thus highly unequally distributed, while in all other advanced democracies it is uncoupled from employment and more equally distributed. It is even possible that mass unemployment – not under present social organization, but in a levelers world – could be profoundly liberating, enabling people to work, individually or cooperatively, on endeavors they value that are not necessarily related to the production of material goods and services. As AI facilitates efficient production, it could also facilitate effective pursuit of these other aims.

In closing, AI is likely to have huge, transformative societal impacts, for good or ill, but present patterns of development and deployment suggest that small "AI for good" movements are likely to be overwhelmed by massive developments that serve concentrated commercial, political, and strategic interests. With such labile technology and such potentially vast impacts, the possibilities for positive

by: Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nick Novelli

transformative change are real, but highly uncertain in their detailed requirements and pathways – and are not being pursued with resources commensurate with their importance, or with the resources directed to systems serving private or rivalrous advantage. With such huge stakes, it is clearly worth pursuing even ill-defined and speculative investigations of how to effectively shift the balance toward the good. We have identified a few possibilities that seem promising, but we fully realize that these are the output of just two days of discussions and are speculative, incomplete, and under-specified. Yet despite all the challenges, further pursuit of these questions, drawing on more breadth of relevant expertise, is a high and urgent priority.

1. David Collingridge, *Social Control of Technology*. New York: St. Martin's Press, 1980.
2. Melvin Kranzberg, "Technology and History: Kranzberg's Laws." *Technology and Culture* 27:3 (544-560), July 1986.
3. See E.A.Parson et al, "Artificial Intelligence in strategic context: an introduction," 2019, at https://aipulse.org; See also S.D.Baum, "Reconciliation between factions focused on near-term and long-term artificial intelligence," *AI and Society* 33(4), 565-572 (2018).
4. See, e.g., Bill Joy, "Why the future doesn't need us," *Wired* April 1, 2000; Samuel Gibbs, "Elon Musk: regulate AI to combat existential threat before it's too late," *The Guardian* 17 July 2017; (plus a great deal of dystopian sci-fi and young-adult fiction).
5. Albert O. Hirschman, *The Passions and the Interests: political arguments for capitalism before its triumph*. Princeton University Press, 1977; Elizabeth Anderson, *Private Government: how employers rule our lives (and why we don't talk about it)*. Princeton University Press, 2017.
6. J.S. Mill, *On Liberty*. John. W. Parker and Son, London: 1859.

7. In this approach – outlining the characteristics of an aspirational set of future conditions, then reasoning through requirements and transitional trajectories to achieve it, we follow the "backcasting" approach pioneered for energy and environmental planning by John Robinson and his colleagues. See, e.g., J.B. Robinson, "Energy backcasting: a proposed method of policy analysis," Energy Policy 10 (1982), 337-344; J.B. Robinson, "Future subjunctive: backcasting as social learning," *Futures* 35 (2003), 839-856.
8. Herbert A. Simon, *The Sciences of the Artificial*. Cambridge: MIT Press, 1969; R.L. Keeney and H. Raiffa, *Decisions with Multiple Objectives*. Wiley: New York, 1976.