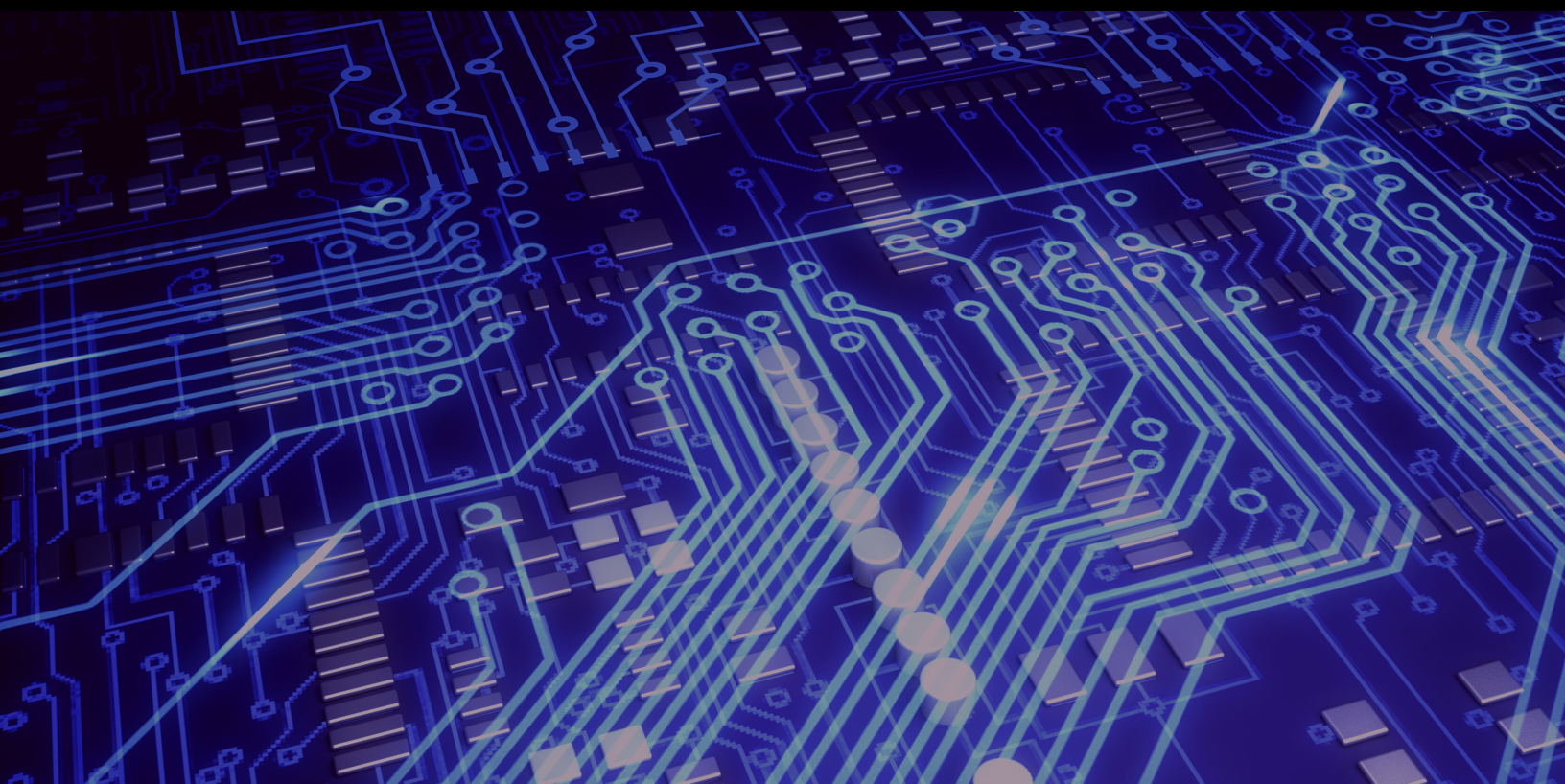
EAST-WEST CENTER

Humane Artificial Intelligence

Working Paper No.03 | October 2020

Ethics and the Risks of Intelligent Technology: The Algorithmic Threat to Freedom of Attention

Peter D. Hershock



Ethics and the Risks of Intelligent Technology: The Algorithmic Threat to Freedom of Attention

Peter D. Hershock, East-West Center

Big data, machine learning and artificial intelligence are transforming the human-technology-world relationship in ways that are complex, recursive, and minimally governed. Financial markets are being transformed by high-speed, algorithmic trading. Workplaces are being transformed by both algorithm-mediated hiring practices and the movement of intelligent machines from factory floors into accounting offices and engineering departments. Machine learning systems are revolutionizing medical and pharmaceutical research. And, governments are turning to big data and artificial intelligence to solve public administration problems, opening pathways to futures in which government for the people might not be conducted primarily by people.

The commercial and political allure of these transformations can hardly be overstated. But, technological transformations are never risk-free. This is especially true of the transformations occurring with the emergence of intelligent technology. Unlike all previous technologies, the informational and computational technologies involved in the now-ongoing Intelligence Revolution are not passive conductors for human intentions and values. They are active and innovative amplifiers of those intentions and values. Intelligent technology thus has the potential not only to scale up the risks of accident and misuse that come with all new tools and technologies, but also to multiply them by the distinctive structural or relational risks that arise when decision environments are refashioned both rapidly and recursively in alignment with uncoordinated and often conflicting values. In short, intelligent technology is liable to create

conditions for an exponential "squaring" of the risks involved in technologically scaling up human intentions and values conflicts.

Considerable attention in media and policy circles has been directed to the far scientific horizon of the Intelligence Revolution and the existential risk to humanity that would be posed by the advent of artificial superintelligence. This is undoubtedly prudent. Were it to occur, this so-called *technological singularity* might well bring about human obsolescence or extinction. But, for reasons I hope to make evident, the technological transformations already underway are at risk of precipitating our ill-prepared arrival at an *ethical singularity*: a point at which the evaluation of value systems assumes infinite value.

Ethics can be variously defined. But at a minimum, ethical reflection involves going beyond the instrumental use of our collective human intelligence to more effectively reach existing aims, and using it instead to discriminate qualitatively among our aims and our means for realizing them. Ethics is the art of human course correction. The ethical singularity toward which humanity is being hastened as machine intelligences tirelessly and recursively amplify human intentions and values is the point at which the opportunity space for further human course correction collapses—a point at which we will have no greater chance of escaping our conflicting human values than light has of escaping the singularity of a cosmological black hole.

Risk, Intelligence, and Technology: Some Conceptual Preliminaries

Risk, intelligence and technology are commonplace words that mean very different things to different people and in different contexts. To help bring into focus the unique risks of intelligent technology, some preliminary reflections on these terms are helpful.

Risk. Risks are not naturally occurring phenomena like apex predators or submerged rocks. Risks are unwanted potentials that cannot be located precisely in either time or space.

Potentials for the presence of apex predators or submerged rocks become risks only if one is considering whether to trek through a predator habitat or to dive into the otherwise inviting waters of a cliff-ringed tropical bay. Risks are virtual dangers. Unlike actual dangers—the presence of a Bengal tiger a dozen yards away—risks are shadows between where we are now and where we would like to be. They are regions of uncertainty that may or may not conceal something that will prevent us from arriving at or attaining what we wish, as we wish. To recognize risks is to discount what we presently value, factoring in the potential costs of pursuing what we want.

Predators on jungle paths and submerged rocks in tropical bays are *not* risks in the absence of intentional, experiential presences that might become prey or injured divers. Although it is possible to accord risks the appearance of material objectivity by assigning them probabilities of occurring and then multiplying these probabilities by the assessed costs of their impacts—thereby generating their so-called expectation value—the virtual existence of risks is irreducibly a function of subjective, *experiential* perspective.[1] The risk of hitting submerged rocks when cliff-diving is not the same for skilled and novice divers. The risk of encountering an apex jungle predator is not the same for a tourist trekker and an indigenous hunter.

The perspectival nature of risk means that risk assessments are neither neutral nor objective. Risk assessments are value-laden. It also means that, until the advent of general artificial intelligence capable of independent value-generation, no artificial or synthetic machine intelligences—including those based on evolutionary algorithms—will be autonomously risk averse. Machine intelligences may conduct themselves in ways that lower risks, but only as long

[1] A brief and usefully general treatment of technological risk is Hansson, 2005. For an extended discussion of the ethics of risk, see Hansson, 2013.

as and to the extent that they have been computationally directed to do so. They will not be capable of factoring new risks into the ways in which they adaptively rewrite their own code or expand the horizons of their conduct. This may not adversely affect their functioning as designed. It will, however, constitute a serious down-the-line risk to humans.

Intelligence. Intelligence is a contested concept, especially when it is qualified as "social" or "emotional," and even more so if qualified as "artificial," "machine" or "synthetic." For present purposes, it is useful to understand intelligence as *adaptive conduct*. This links intelligence to both learning and goal orientation, but not necessarily to self-awareness or subjective experience. "Adaptation" implies both permutation and persistence in a changing environment, while "conduct"—which comes from the Latin *conducere* or to "bring together"—stresses relationality or blended action. Adaptive conduct thus involves at least minimal interdependence of the actor and acted-upon, a degree of causal mutuality. Put somewhat differently, intelligence is always embedded within and shaped in relation to a dynamic environment of actionable possibilities.

Adaptation is never certain. Intelligence, broadly construed, thus entails sustained creative anticipation in conditions of uncertainty. Intelligence is not to be confused, then, with knowledge or even with the capacity to learn, though learning depends on it. Intelligence operates at the margins of knowledge. Every living thing embodies a complexly rewarded history of adaptive successes: intelligence demonstrated across evolutionary time scales. In addition to such slow, evolutionary intelligence, some living beings, like domesticated animals, are capable of "fast" adaptations to environments full of uncertainties with which they have no prior experience. Human beings, perhaps uniquely, actively seek and create risk-saturated

environments in which to manifest and explore new types of intelligence—for example, wars, sporting events, romantic relationships, and market competitions.

It is open to debate whether all living things have at least minimal subjective awareness and hence the potential for perceiving and adapting in light of risks. But, organic intelligence is mortal. Manifesting in spaces of sustained interdependence between relatively autonomous organisms and their environments—organic intelligence is honed by adaptive confrontations with life-ending eventualities. This orientation toward survival implies perspective-bound intentionality and perceptions of value. Seeking, evading and abiding are basic forms of organic intent. Attraction, aversion and neutrality are basic organic values. Organic intelligence thus entails adaptive values-differentiation—generating increasingly refined and qualified assessments of environmental affordances in terms of what not only *can*, but *should* be pursued, avoided, or ignored in conditions of uncertainty. And, in the face of new dangers, organic adaptive success entails at least implicit considerations of risk.

Today's state-of-the-art computational intelligences are extraordinarily fast. They are capable, in mere seconds or minutes, of thoroughly exploring an environment of actionable possibilities—for example, a database of millions of photographs—that would require years for humans to explore. But their exploratory perspective is both predetermined and extraordinarily narrow. If these inorganic intelligences are capable of something like implicit considerations of risk, these risks are purely epistemic. Even for evolutionary deep learning algorithms that rewrite their own computational DNA in response to environmental feedback, the only "risk" implicit to their conduct is that of being wrong.[2] Not getting things right is, of course, valuable feedback.

[2] This limitation may be a function of the current state-of-the-art of machine intelligence or it may be intrinsic to any computational system. See: Adam Safron, Frontiers in Artificial Intelligence, 2020. "An Integrated World Modeling Theory (IWMT) of Consciousness: Combining Integrated Information and Global Neuronal Workspace Theories

Making mistakes is an excellent way to learn. But, if unconstrained by perceptions of risks other than that of being wrong, learning through mistakes can easily amount to a practice of naively wronging those "brought together" in the causal mutuality of adaptive conduct.

Technology. Technologies are often identified with specific tools. Information and communications technology, for example, is often identified with smartphones and computers. But technologies are not things. Technologies are emergent, relational systems of material and conceptual practices that embody and deploy both strategic and normative values, qualitatively transforming the ways we relate to the world around us and with one another.[3] Technologies not only shape *how* we do things—like communicating and moving from place to place—they also shape *whether* and *why* we do so. Technologies are environmental systems for scaling up and structuring human intentions in ways that, over time, also change the meaning of what is intended.[4]

Tools extend, augment or otherwise alter our capacities for carrying out particular kinds of work. Tools have task-specific utilities and are thus aptly evaluated in terms of how well they enable us to accomplish what we want. Tool designs constrain, but do not specify, how tools are employed, and no matter how popular or seemingly "essential" certain tools may be for carrying out specific kinds of work, we can always opt out of using them. Tools, even those as large as electricity grids, are localizable, and with respect to them, we can always exercise "exit rights."

with the Free Energy Principle and Active Inference Framework; towards solving the Hard problem and characterizing agentic causation."

[3] For a short discussion of this distinction with reference to information and communications technology, see Hershock, 2017; for a more extensive discussion, see Hershock, 1999.

[4] One implication of this way of understanding technology is that, although we tend to associate technology with transforming relations with the material world, societal institutions—e.g., investment and exchange systems and marital institutions—function technologically to scale and structure intentions and actions in ways that dramatically shape relational dynamics. What distinguishes such "immaterially" focused technologies is their primary emphasis on shaping the human dimension of the human-technology-world relationship.

This is not true for technologies. Unlike tools, technologies are not localizable. In actuality, we do not *build* or *use* technologies, we *participate* in them. Technologies emerge from and inform/structure our conduct in the same way that natural ecosystems emerge from and dynamically inform/structure species relationships, or the ways that urban environments and their complex systems of infrastructure simultaneously open new possibilities for action and habituate us to engaging in certain kinds of actions for specific kinds of purposes. Technologies are relational media within and through which we exercise and extend the reach of our intelligence in pursuing what we value. That is, technologies are *marriages of structure and agency* that cannot be adequately evaluated solely in terms of task-specific utilities and from which we cannot effectively exercise exit rights. Technologies can only be fully evaluated *ethically,* in terms of how they qualitatively affect human-human and human-world relational dynamics.

The Relational Complexity of Technological Risk

The technological scaling and structuring of human intentions and values is inherently risky. In part, this is because technologies are historical and expansive. They persist over time in ways that generate ever-further opportunities for extending the reach of human intentions, bringing more and more of the natural and social worlds within the purview of technological intervention. As technologies saturate new natural and social environments, however, they open prospects for scaling up both mutually-reinforcing and competing values and intentions.

The technological advancements brought about by the invention of writing systems enabled scaling communicative intent far beyond the sonic limits of speech to reach "audiences" thousands of miles and years distant. The advancements brought about by the invention of steam-powered engines made transportation sufficiently fast, cheap and reliable to enable

commercial intentions to operate at a truly global scale. Yet, print and later broadcast communications technology also opened communicative pathways around the mirror neuron system that mediates empathetic sociality in face-to-face communication, thus expanding the risk of successful, prejudice- and violence-inciting mass indoctrination. Transcontinental and transoceanic shipping and travel induced a shift from local subsistence agriculture to global commercial agribusiness that spawned both new food security risks and increasing risks of diseases and pests spreading into immunologically unprepared populations.

The technological origins of eventualities like these are often dismissed by labeling them accidental or unintended. But this is a kind of philosophical sleight-of-hand—a redirection of critical attention from technologies to tools. As is often stated by gun rights advocates, "guns don't kill, people do." In fact, however, weapons technology scales and structures human intentions to inflict harm while making apparent the value of doing so from as great a distance as possible. Weapons technology thus transforms the meanings and stakes of both attack and defense, reconfiguring decision-making environments as disparate as homes riven by domestic violence and geopolitical orders riven by fears of hegemonic intent.

If attention is diverted from technologies to tools, it seems reasonable to address the risk of children accidentally harming themselves while playing with loaded guns or that of crimes being committed with stolen weapons, by ensuring that all guns are built with biometric identification systems that keep them from being used by anyone other than their registered owners and those to whom use rights have been formally extended. The risk that social media platforms might be used to foment racist sentiments or to identify and indoctrinate potential terrorist recruits might be addressed technically through a rigorously designed system of human or algorithmic gatekeepers.

But, such tool fixes, no matter how extensive or ingenious, are impotent with respect to the risks involved in reconfiguring human-human and human-world interdependencies. Technical fixes like building better and safer tools place constraints on acceptable utility. They address the final causal phase when an agent and a tool-enabled action inflict harm. But, as noted earlier, technologies are values-infused decision-making environments that are structured by and in turn restructure human agency and values, and the human-technology-world relationship is one of complex, network causalities. Technological harms are not *pre-identifiable events*; they are *emergent phenomena* that grow recursively over time, becoming part of the relational fabric of practical decision-making.

Mass shootings are events made possible by guns. But while a gun is necessary condition for a mass shooting, the shooting itself is merely the culmination of a painfully distorted pattern of human relationality that has made inflicting mortal harm at a distance appear to be a reasonable means of extrication from it. Mass shootings and criminal gun violence do not occur in psychic, social, economic or political vacuums. The *readiness* to inflict harm at a distance is a risk of weapons technology. Understanding and addressing technological risk is not a process of anticipating and guarding against future harmful events; it is a process of *anticipating dynamic relational patterns* that become harmful as they are scaled up to become part of the environments in and through which we express what it means to be human.

The Distinctive Risks of Intelligent Technology

We are only beginning to understand the distinctive risks of intelligent technology. It was only over the first half of the 2010s, as global data production skyrocketed along with smartphone use, 24/7 wireless connectivity and a rapidly burgeoning internet of things, that algorithmic machine intelligences began migrating out of research labs. Tasked with adaptively

responding to a remarkably wide array of real-world problems and environments, the results were stunning. In just a handful of years, after many decades of slow and incremental progress, machine learning systems were able to equal or surpass human levels of image and speech recognition. They managed to teach themselves how to play unbeatable *go*, to perform at professional levels in multiplayer real-time strategy games, and to accurately predict cancer treatment discoveries by combing through thousands of decade-old research papers.

For the technologically hopeful, this "Cambrian explosion" of data-nourished machine intelligences marked nothing less than the beginning of a 4th Industrial Revolution or Second Machine Age that would free humanity forever from material want (see, e.g., Brynjolfsson and McAfee, 2014). Aside from the expected risks that industry-transforming creative disruption would pose to entrenched interests, the greatest risk ahead was the opportunity-loss risks of failing to act early and decisively in building out AI capabilities or of being held back by stiff regulatory limits on the free play of technical, scientific or commercial imaginations.

For the technologically cautious, the evolutionary explosion of machine intelligences made suddenly real the venerable science-fiction trope of an emergent artificial superintelligence directed toward other-than-human ends and interests (see, e.g., Bostrom, 2014). The greatest risk ahead was not just the accelerated dehumanization of technological entanglement that had been worried about since the profit logic of the 2nd Industrial Revolution began subordinating humans to machines on factory floors. It was nothing less than human obsolescence or extinction.

Over the last five years, as the daily life interventions of data-nourished algorithmic intelligences have become increasingly pervasive, a new and topographically complex terrain of immediate and near-term digital dangers has appeared between the opportunity-loss and existential-risk horizons of the technological optimists and pessimists. These dangers include the

risks of unchecked, data-driven abuses of private/corporate power (Zuboff, 2019; Wu, 2016); the risks of inapt (and inept) design and application of AI-enabled tools by both corporations and governments (Eubanks, 2019; O'Neil, 2017); the risks generated by autonomous weapons systems capable of executing military directives with inhuman speed (Etzioni and Etzioni, 2017); and the personal, corporate, and national risks posed by the infinite attack surface resulting from ubiquitous thing-to-thing (and not just person-to-person) digital connectivity (Schneier, 2018).

Ethics in Anticipation of Technological Risk: An Incomplete Effort

In apparent acknowledgement of and response to this emerging risk environment, a vibrant global cottage industry has developed around crafting artificial intelligence and data ethics principles and guidelines.[5] The common features of the many dozens of ethics principles and guidelines that are now in circulation are telling.

To begin with, while these sets of ethics principles and guidelines all nod toward the existential risk of intelligent technology and universally affirm commitments to AI that is human-centered and aligned with human values, they have generally done so without defining what is meant by "human centered" and without specifying which human values are to be included and with what priority. Given that ethics—the art of human course correction—often involves choosing among competing values, this is a significant shortcoming.

In addition, there has been a notably universal stress on the importance of developing intelligent technology that is accountable, explainable, robust, safe, fair and privacy-protecting. These are also generally presented as uncontroversial commitments. No explicit mention is made of the fact that known and substantial differences exist, for example, in how privacy is

[5] For a critical review of AI ethics principles and guidelines, see Hagendorff, 2019.

understood and respected in various cultures or in what constitute reasonable safety standards, even for such clearly dangerous applications as autonomous weapons.

Finally and most importantly, however, with very few exceptions to date, AI principles and guidelines are strikingly silent about the potential societal costs of intelligent technology—for instance, the potentials for mass unemployment and underemployment, for negative impacts on social cohesion, for entrenching existing patterns of inequality, or for carrying out social or political engineering. Given the widespread expression of concerns about technological unemployment, an arms race in autonomous weapons systems, and the possibility of AI-driven social and political engineering, this silence signals a division of labor according to which the responsibility for addressing social costs is tacitly relegated to political and institutional actors, with debate centering on appropriate public policy rather than ethics.

In sum, what we find expressed in the majority of AI ethics principles and guidelines to date is a framing of intelligent technology as ethically neutral, with critical attention directed toward the human actors engaged in designing and using the "smart tools" that the technology is making possible. Concerns about accountability, explainability, safety and data fairness are, of course, well-warranted and ensuring that they are taken seriously is unquestionably central to the practice of ethical engineering. But, these are essentially engineering goals that are amenable to being defined and achieved computationally. They are *technically* tractable.

Algorithmic bias resulting from skewed training data in otherwise benign uses of artificial intelligence can be fixed by using more complete data sets and by reducing the use of stand-in or proxy data in predictive analytics—using zip codes to presort loan applications, for example, in lieu of complex, causally-relevant data regarding a loan applicant's loan payment behavior. The use of machine learning systems to identify people—like single mothers, recent

veterans, immigrants, and the poor—who are likely vulnerable to offers for educational loan "assistance" from predatory, for-profit colleges can be subject to stricter regulation enforced with the help of proactive, industry-wide algorithmic audits.

Technical approaches like these are well-suited to mitigating the risk of harms directly resulting from *accidents of design* and from *misuse by design*, and this is important work. But these are hardly the only risks of algorithms being deployed, for instance, to rate and rank people for school admission, employment, and loan applications. How might distributions of income, wealth and opportunity be restructured if behavioral prediction—which is central to the vitality of the digital network economy—becomes ubiquitous? How consistent is this likely to be with democratic values? How might it affect social cohesion and resilience? Are the efficiency gains that "smart" predictive services afford schools, businesses and government programs valuable enough to offset the personal and societal harms that might result if people are algorithmically held hostage by their past or by the pasts of their family members and friends?

These are not technical questions; they are ethical. Determining whether a given technology *can* be deployed widely, accountably, safely and fairly in society does *not* provide answers to questions about whether it *should* be deployed. The Intelligence Revolution promises abilities to scale up intentions to treat disease, to shrink carbon footprints, to connect people with shared interests and concerns, and to make all information available to anyone with an internet connection. Smart tools and smart services may exhibit only very narrow intelligence, but the deep learning methods and artificial neural network logic underlying them are extraordinarily flexible. The scope of their applications is virtually unlimited. And that is one of the perils of the Intelligence Revolution. These same computational tools can be used to design killer viruses, to hack into power grids, to foment populist anger, and to carry out misinformation campaigns.

These are serious risks to weigh against the anticipated benefits of intelligent technology. Focusing exclusively on the technically-tractable, accidental and misuse risks of intelligent technology, however, can easily amount to simply "green lighting" its pervasion of society and turning a blind ethical eye to its more substantial structural risks.[6]

The fact that many of the most prominent AI ethics principles and guidelines have been crafted by teams funded by corporations and governments with vested and competitive interests in intelligent technology raises worthwhile questions about whether this "green lighting" is entirely inadvertent.[7] But, as tempting and societally important as this line of questioning might be, it unfortunately turns ethical attention back to the possibility of harms brought about by "bad apples" among the corporate and state actors funding the growth of intelligent technology to further their own interests. Slapping a few hands, however, will not change the "green light" that has been given thus far to "smart" transformation of the human-technology-world relationship.

As evidenced in nearly all AI ethics principles and guidelines to date, one impediment to changing the light from green to yellow or red is the traditional ethical emphasis on individual (human) agents, which tends to direct attention to the final causal moment in the experience of harm and away from the complex, nonlinear causal dynamics that characterize the human-technology-world relationship. That is, it directs attention away from the ways in which technologies change the nature of the environments within which people act, interact, collaborate and compete, as well as the incentives for doing so. In short, exclusive reliance on ethics framed around the centrality of individual agents, actions, and patients facilitates continued blindness to

[6] On the accidental, misuse and structural risks of AI, see Zwetsloot and Dafoe, 2019.
[7] For a very disturbing expose along these lines, see: https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/?comments=1

the risks of intelligent *technology* in favor of identifying and mitigating the risks of smart, computational *tools*.

Risking Ethical Singularity

Intelligent technology actively and adaptively scales and structures human intention in ways that are both inhumanly fast and free of the organic limitation of being intrinsically risk-averse. If there are conflicts among the values that inform the human intentions, these conflicts will also be scaled with inhuman speed and acuity. That is already a significant risk. Granted that ethics is the art of course correction in response to conflicts of values, aims and interests, perhaps that is a risk worth taking. Optimistically, intelligent technology might serve as an accelerator of experienced needs for ethical deliberation and course correction, and thus as an amplifier of efforts to engage systematically in collaboratively evaluating value systems. This could have very positive consequences.

But technologies emerge out of and recursively reinforce dynamic structures of human-human and human-world relationships. These relationships are fundamentally epistemic and strategic—both knowledge-seeking and knowledge-enacting. They are necessarily also cultural, social, economic and political. Intelligent technology can most generally and neutrally be described as an emergent system of material and conceptual practices for gathering and computationally sieving data—the direct and indirect traces of intelligent human conduct—to resource solutions to human problems. Motives vary among those who are responsible for designing and building the computational factories and smart tools that are involved in transforming data into solutions, but they are clearly not exclusively altruistic. The same is true of the motives of the national and regional governments that are sanctioning and investing most heavily in intelligent technology—notably, the U.S., China and the EU. These motives matter.

To anticipate how these motives may be affecting the emergence of intelligent technology and its associated risks, it is useful to "follow the money." As flows of investment capital make clear, the future of the global economy is no longer to be found in the manufacturing, mining, energy, financial or retail sectors. The seven largest companies in the world by market capitalization in 2019 were all AI, e-commerce and connectivity giants: Apple, Amazon, Alphabet, Microsoft, Facebook, Tencent and Alibaba. In spring 2020, the market value of Netflix surpassed that of Exxon Mobile. It is now the systematic attraction and exploitation of attention that drives global circulations of goods and services.

In this new attention economy, revenue growth is becoming ever more closely proportional to the effectiveness of machine agencies in discerning what we value and predicting what we can be induced to want. It is an economy in which major platform, social media, and internet service providers are empowered to profit from algorithmically tailoring our increasingly connection-defined human experience, and in which the holy grail is not free energy or cheap labor, but *total attention-share*.[8] Through participating in this new global attention economy, in exchange for their attention and the data carried with it, consumers are provided with individually-tailored and seemingly infinite arrays of choices for satisfying desires for tangible goods and services, for the intangible experiences afforded by film and music, and for social connection. The ultimate commercial goal is the frictionless and ceaseless conversion of consumer attention capital into profitably satisfied desires.

Sustained attention, however, is the single indispensable resource required to engage intentionally in relational transformation. The nominally free, infinitely personalized

[8] For fuller explorations of the attention economy, see Hershock, 2012 (pp. 134-147) and Chapter 3 of Hershock, 2021 (forthcoming).

connectivity that we receive in exchange of our attention and data exhaust is not, in fact, either costless or value-neutral. The epistemological powers of intelligent technology are also ontological powers. They are powers to sculpt who we become, not through acts of coercion, but through the systematic prediction and satisfaction of desires and the individualization of unevenly beneficial decision-making environments.

The resulting concentration of corporate power, and the funneling up of wealth and opportunity that accompanies it, is cause enough for serious global worry. The risks of misuse are immense. But the deeper concern is that the predictive and responsive powers evolving along the commercial interfaces of human and machine intelligence are transforming the structure of human experience from the inside out. As small numbers of global elites broker what amount to arranged marriages of the *attention economy* and the *surveillance state*, each intent on maximizing the competitive advantages afforded by gains in intelligent technology, the interplay of human attention and intention in itself is at risk of being transformed into a primary "theater of action" in what ultimately amounts to the colonization of consciousness itself.

The Ethical Risk of Algorithmically Exploiting Attention and Desire

Considerable evidence now exists that the basic work of minds involved in adaptive conduct is not that of discovering *what* things are, but rather discerning *how* things are changing (Clark, 2016)—an effort to *anticipate* patterns of change in spaces of uncertainty. "What comes next?" may be the most basic question of all. But, it is not an inherently simple one. The intelligent work undertaken by machine learning systems is essentially predictive processing. Asking what comes next is about *forecasting*, the power of which is proportional to the accuracy of the causal maps that can be built about how things—and, of course, we humans and our thoughts, desires and actions—are changing.

For organic intelligences, anticipating what comes next is also about preparedness or embodied *readiness*—an ongoing, attentive process of intuiting relational scenarios or assemblages of likelihoods, and preparing responsively in light of them. What matters most in this case is not accuracy, but adaptability or responsive immediacy and fitness. This implies prospective, evaluative concern about the interplay of what *could* and *should* occur. Intelligence here consists in *affecting* relational dynamics: determining what matters most in seeking and securing what matters most.

The algorithmic attraction and exploitation of human attention to shape human intentions and behaviors has the potential to short circuit this process. Attention qualitatively shapes anticipation, endowing it with dimensions of breadth and depth. This can consist in perceptual directedness toward registering relevancies and change dynamics in actionable environments, both internal to and external to the body, or it can consist in enactive directedness toward reconfiguring change dynamics in those environments. Intentions mediate the translation of attention-shaped anticipation into action.

As noted earlier, one of the liabilities of algorithm-based predictive analytics is their capacity for holding people hostage to their own past and the pasts of those with whom they are related by birth, voluntary affiliation, and by accident of geographic/social location. The new, digital attention economy and its algorithm-engineered maximization of attention capture and desire turnover is liable to holding people hostage to their past likes and dislikes, their patterns of preference and aversion. And, it is liable to doing so in ways that are both critically opaque to those affected and conducive to attention resource deficits severe enough to place at risk capacities for and commitments to anticipatory course correction or prospective concern about what *should* occur. Diligently and deftly enacted and reinforced by machine intelligences,

commercial and geopolitical ambitions have the potential to place at risk the *freedom of attention* without which ethically creative course correction becomes impossible.

The probability of this risk might seem small. After all, capturing, holding and exploiting attention is nothing new. "Attention merchants" have been trading systematically in human attention and converting it into revenue for at least two hundred years (Wu, 2016). Magicians, musicians, religious leaders and young people in search of mates have since time immemorial attracted and directed attention for reasons ranging from the utterly instrumental to the transcendental. The algorithmic attraction and manipulation of attention may be taking place at entirely unprecedented scales and speeds, and with unprecedented precision. But, participation in intelligent technology via the infrastructure that allows consumers/users of digital connectivity and commerce platforms to function simultaneously as producers of algorithmic intelligence training data is not total. Yet, even if total attention share remains a goal rather than an achieved reality, this shortfall is more than compensated for by the fact that the digital attention economy is embedded in and dynamically informs the expository society (Harcourt, 2015). It is an economy rooted in ambient, inhumanly deft amplifications and exploitations of desire.

Desire is an imaginative kin of anticipation. It begins as sustained attention to a "present absence"—an anticipatory yearning for completion. When it is strong enough, however, desire becomes attention-arresting. In contrast with anticipation in conditions of uncertainty, desire is certainty of want or lack. To be wholly caught up in desire is to cease predicting and readying, and thus it is not inaccurate to say that one can lose one's mind to desire. When paired with an intention to act in ways that will lead to attaining what one wants, desire can helpfully sharpen

the focus of intelligence. But, when desire becomes an all-encompassing seeking of gratification it dulls the anticipatory responsive edge of intelligence; action becomes compulsive.[9]

The technological scaling of desire and its digital satisfaction has the potential to facilitate the sculpting of desire-defined individuals whose attention is only nominally their own and for whom the ethical art of course correction through the resolution of values conflicts is happily left to absent others in favor of continuing to exercise unconstrained freedoms of connective and experiential choice. But, without freedom of attention, there is no freedom of intention. Without freedom of attention and intention, action tends to devolve from intelligent, anticipation-qualified response into either habit-determined or instinctual reaction. Technology that innovatively scales up (perhaps conflicting) intentions to exploit attention and desire runs the risk of creating conditions for an ethical singularity—a point at which capacities-for and commitments-to collaboratively evaluate aims and values collapse as the structural means by which interests are met and desires satisfied become perfectly automated.

Mitigating technological risk involves fundamentally ethical labor. To avoid arrival at a technologically induced ethical singularity and address the complexity of the risks associated with intelligent technology will require more that the rigorous application of one or another system of ethics. All ethical systems have blind spots. To anticipate and mitigate the risks of accidents of design and the risks of misuse by design associated with intelligent technology, it may suffice to develop robust codes of professional conduct and globally-operative legal institutions. To anticipate and mitigate the structural risks of intelligent technology, however,

[9] In Buddhist reflection on desire, this distinction is captured conceptually. To be subject to a craving for or clinging to what one does not have is unwise. Desire the *compels* is *taṇhā*. But the action-intending desire involved in striving to be more virtuosically responsive or to bring about more liberating relational dynamics can become the root of wisdom. Desire that *impels* practice is *chanda.*

will require sustained international, intercultural and intergenerational ethical deliberation and improvisation. We are not in need of a new species of ethics. We are in need of a resiliently diverse ethical ecosystem.

Works Cited

Beck, Ulrich (1999). *World Risk Society*. London: Polity Press

Bostrom, Nick (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford, UK: Oxford University Press

Brynjolfsson, Erik and Andrew McAfee (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York: W.W. Norton

Clark, Andy (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*, Oxford University Press, 2016

Etzioni, Amitai and Oren Etzioni (2017). "Pros and Cons of Autonomous Weapons Systems," Military Review, May-June 2017, available online at: https://www.armyupress.army.mil/Portals/7/military-review/Archives/English/pros-and-cons-of-autonomous-weapons-systems.pdf

Eubanks, Virginia (2019). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: Picador.

Hansson, Sven Ove (2005). "The Epistemology of Technological Risk," Techné 9:2, pp. 68-80

_____ (2013). *Ethics of Risks: Ethical Analysis in an Uncertain World*, New York: Palgrave MacMillan

Harcourt, Bernard (2015). *Exposed: Desire and Disobedience in the Digital Age*, Cambridge, MA: Harvard University Press.

Hershock, Peter D. (1999). *Reinventing the Wheel: A Buddhist Response to the Information Age*, Albany, NY: State University of New York Press

Hershock, Peter D. (2017). "Ironies of Interdependence: Some Reflections on Information, Communication, Technology and Equity in Contemporary Global Context," in China Media Research, October 2017.

O'Neil, Cathy (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Broadway Books

Piketty, Thomas (2017). *Capital in the 21st Century*, translated by Arthur Goldhammer. Cambridge, MA: Belknap Press

Schneier, Bruce (2018). *Click Here to Kill Everybody: Security and Survival in a Hyper-Connected World*, New York: W.W. Norton

Tononi, Giulio (2012). *Phi: A Voyage from the Brain to the Soul*. New York: Pantheon

Wu, Tim (2016). *The Attention Merchants: The Epic Scramble to Get Inside Our Heads*. New York: Knopf

Zuboff, Shoshanna (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs.

Zwetsloot, Remco and Allan Dafoe (2019). "Thinking about Risks from AI: Accidents, Misuse and Structure," *Lawfare*, February 11, 2019. www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure

Humane Artificial Intelligence

Humane Artificial Intelligence is an intercultural, intergenerational, and interdisciplinary initiative of the East-West Center (EWC) to engage the societal challenges and opportunities that are emerging with advances in artificial intelligence, robotics, machine learning and big data. The purposes of the initiative are: to establish an ongoing forum for exploring differences in understanding the goal of aligning artificial intelligence with human values and societal wellbeing; to foster globally-shared commitments to equitable and humane artificial intelligence; and, to engage emerging perspectives on the evolving interplay of people, artificial intelligence, and related technologies.

The views expressed in this article are those of the authors and do not represent the views of the East-West Center.

EAST-WEST CENTER