Preface

The collection begins with three critical appraisals of the potential for high technology and intelligent machine solutions to resolve large-scale social, economic and environmental problems. Our first chapter is from Rick Searle, our co-editor. This chapter leads for two reasons. One, it is easily amplified by succeeding contributions as it touches on many of the themes running through the text as a whole. Rick confronts social-political problems arising with the increased automation and indeed out-sourcing of social-political life through intelligent machine technologies. From the reliance on smart-phones to locate friends and to coordinate political movements, to the automation of surveillance and thus to the dangers of technologically enforced neo-feudalism, Rick's purpose is profound, and the case that he presents disturbing. Writing that "the most potent critique against digital teleology is that it results in a kind of moral atrophy where human beings become the puppets of a world they have themselves dreamed up", Rick points out that some researchers pursuing artificial intelligence at once believe that their efforts will result in artifacts "that will then go on to destroy humanity." The relationship with the introductory pages of this preface is clear enough. Rick confronts the issue with candor constituting the second reason for this entry's lead role in the text:

> *Eschatological determinism of such a sort has more in common with religious fundamentalism than it does science, and raises serious doubts over the goals of those who have heretofore not faced major ethical or political design constraints when building the Internet or pursuing artificial intelligence. … The fact of the matter is that technological development is less about human survival this century let alone the "destiny" of life in the universe than it is about political and economic power as it is manifest right now.*

Our second chapter is from Chris Bateman, a philosopher of games, as well as a programmer and leading designer of narrative structures for games and virtual environments. Chris embraces the immediacy of our situation as established in the first chapter, setting out from the following dilemma: "Having decided that we *can* make anything, we must now ask the vital question: *should* we?" He weaves discussion on two burgeoning fields of intelligent machine technologies, automated transport and smart weaponry, in terms of Kantian, utilitarian and virtue based approaches to ethics of technology finding at the heart of these industries a deep yet neglected moral disconnect between stated aims and actual results, with the growing dependency on both the technologies and the distancing that they afford from unpleasant consequences constituting a profound symptom of what he terms "*cyberfetish*".

Through the glaring moral angle of cyberfetishism, Chris confronts social-political issues of a form and in a frame reinforcing Rick's first chapter. Chris' analysis draws out the co-dependent nature of the human propensity to fall under what the first chapter covers under the heading of "algocracy", the rule of automation through algorithm, with human beings ultimately molded in service to the medium of their own ignoble dependency. Chris' cyberfetish drags into the open the essence of technology, that it is habit incarnate, not only addictive but an externalization and artifactualization of dependency, with the technological world of our making directly affording the exercise of recently acquired tastes and supporting habits. Should one generation's habits prefigure the freedom of future generations to embody different habits, healthier habits, as these prior habits are entrenched in layers of technological dependence? Through his discussion, Chris' question thus becomes not so much what are we to build, but for what world of affordances do we wish to be held accountable.

Our third chapter continues this critical revision of the promise of smart-machines, digging into the problem of environmental sustainability from the perspective of sustainability scientist and civil engineer Nak Young Seong with editor and co-author Jared Gassen. On this view, history demonstrates that so-called technological "solutions" to social-political problems are not really solutions at all, and rather evidence shows them to be problems in disguise. New technologies require new materials, new industries, new logistic paths, and create new - sometimes unknown or unsuspected – pollutants and other problems arising from those. The authors take on controversial issues, such as the use of water resources expended to simply cool the massive NSA data collection center in Utah, and argue that the expense to future generations in the form of irremediable environmental damage cannot be justified. Rather than technological "solutions" causing only further technological problems in need of solution, the authors recommend that we step back and reassess our options in light of new ways of understanding sustainability and natural resources. Seong and Gassen advise for the employment of "low-tech" sustainable and effectively non-polluting energy resources like trees as responsible bases for future economies rather than hanging the hopes of humanity on inherently deadly and ultimately uncontrollable technologies like nuclear power only to require robotic laborers to eventually brave an irretrievably polluted and deadly toxic natural environment.

Where the third chapter hints at the evolutionary significance in any project pursuing a machine-leveraged break from the womb of human evolution, the fourth chapter penned by powerhouse Luis Moniz Pereira and Fernando Cardoso demonstrates how these natural forces may have shaped human moral capacities in the first place. From an evolutionary psychological perspective in which counterfactual reasoning arises alongside predictive capacities associated with capacities for belief revision in light of especially social information sourced to other members of an "ethical association" of moral agents, Cardoso and Pereira are able to illustrate how communication of moral commitment helps to strengthen in-group cooperation, thereby shedding light on some of the themes raised so far in this Preface as these communications become at least co-determinative of satisficing goal conditions going forward.

Their evolutionary psychological approach to modeling morally significant decision spaces is deceptively simple, as it plausibly accounts for the evolutionary emergence of group-level behavior patterns associated with moral and ethical abilities and institutions, respectively. The models are built from Prospective Logic Programming (PLP), and the authors review their program in some detail. Details aside, PLP "supports the view that autonomous agents are those capable of anticipating and reasoning about hypothetical future scenarios" with prospective cognition "essential for proactive agents working with partial information in dynamically changing environments." This describes us in our collective situation today. Cardoso and Pereira's confessedly "limited" agents in equally limited decision spaces regardless "do illustrate a rudimentary sort of reflective equilibrium over possible ends" and, with these consequences known, "meta-reasoning techniques are applied to weigh the partial scenarios" exactly as we weigh out possible ends for pursuing this or that line of scientific research, for example. Thus understood, the real promise of this approach is to be able to monitor the influence of different psychologically realistic pro-social characteristics on the emergence of ethical institutions within increasingly large populations of complementary agents, thus providing a medium for the simulation of similar problems in more complex decision environments and with more robust cognitive agents in the future. Luis and Fernando conclude by pointing to future work in the simulation of imitation, deception, and emulation at the agent level, and thusly they expect to refine our understanding of morality as an evolutionarily emergent property, writing confidently that "future work further integrating moral philosophy with programming will establish necessary logical supports to complete the task" and that "although the processes within us are complex, their complexity is not inaccessible" to properly configured computational models thereof. Models like these are great tools for policy-makers and social engineers going forward, and a good deal of the general project in intelligent machines depends on the value of these kinds of simulations in delivering on some of their anticipated promise in helping to solve coordination problems in realizing a healthy, organic and flourishing yet technologically rich world.

Where Fernando and Luis offer an illustrative logic of human morality and the growth of human ethical institutions through the enactment thereof, our fifth chapter develops a mirror on morality also from an evolutionary foundation but here through the smart manipulation of "big data". Rafal Rzepka and Kenji Araki of the Graduate School of Information Science and Technology at Hokkaido University detail the use of text-mining techniques to build a portrait of contextually dependent morally significant terms as they are used by human beings in everyday (online) life, and from this battery of information deliver representative judgments in similar contexts. They adopt a bottom-up intuitionist perspective, allowing a portrait of morality to arise from everyday moral language rather than have episodes cherry-picked to suit model-theoretical preconceptions of moral conduct in order to capture the emotional motivations of moral action, "empathy", writing that "evolution equipped us with emotional reactions that were originally meant for survival, and then to flourish as societies" at least in part through the expression of such feelings. One advantage of this approach is that they are able to "take advantage of the fact that, for the time being, the only constantly growing data that is relatively easy to process is textual" and to use this text-based, symbolic information to reliably gauge moral judgments elicited through human moral emotional mechanisms. Unlike that of the previous chapter, their approach does not involve the reduction to fairy-tale terms of complex moral dilemmas, and rather avoids "sophisticated algorithms" in order to "concentrate on automatically collecting and analyzing descriptions of human behaviors" during "everyday life situations available to current or near future devices with natural language processing capabilities." Thusly, Rafal and Kenji provide both a portrait of morality as actually expressed, as a sort of mirror on a "fourth-person" moral construct, as well as an approach to data-set modeling useful in evaluating expressed moral judgments of current and future first-personal agents relative context.

The potential for moral web-bots auto-texting to influence online discourse is a scary prospect, and quite near given the fifth chapter. In our sixth chapter, Melanie Swan looks behind the moral responses, into the minds of the respondents, themselves. Melanie confronts the issue of "reality" as a public, ethical sphere, open to private manipulation, noting that this only becomes an issue as technologies empower the direct manipulation of everyday perception and action, with her focus being the dedicated industry of perception management through the physical integration of human perceivers with smart perception management nano-machines. She speculates over possible "killer applications" for perception management technologies proffered by "nanocognition", including memory management services which nullify painful memories by destroying neural assemblies encoding them, "bias reduction" as a means for physically correcting for error-prone prejudice, and even the direct sharing of point-of-view experience, suggesting possibilities for conflict resolution. Possible applications like these raise issues of "neural privacy" – as neural structures embody information, how should this information be dealt with, perhaps as a commodity or as a matter of state censorship and law enforcement? Tools such as those under development by the authors of the previous, fifth chapter, may well serve as means for the monitoring of large-scale social engineering projects employing technologies such as those under Melanie's scope, perhaps eventually prior to the literal expression of moral emotion altogether. Her chapter also resonates with the deep themes that had established themselves in the collection thusfar, recalling for instance Chris Bateman's "cyberfetish" from the second chapter, and with her treatment of technological co-dependency feeding equally well Seong and Gassen's contribution. For example, Melanie's discussion allows us to ask if the best use of natural resources is in the development of technologies helping us to perceive only pleasant things, or permitted things, or if these resources are best directed to other ends. It may well be a symptom of a collective human cyberfetish to expect intelligent machines to make all of our painful memories go away, and to fashion new ones only the way that we would want them beforehand.

Our seventh chapter continues the assay of ethical implications due transformative human integration with intelligent machine technologies. Focusing on human enhancement in competitive – especially business - environments in terms of the "biopolitics" of human-machine integration, the author, Ben Tran, asks "whether enhancement technologies will actually make our lives happier". Ben delineates ethics of human enhancement issues according to an "expanding circle" of individual, professional and societal levels of organization. Defining "*societal concerns"* as "the broad interests of society, which may be frustrated by the adoption of human enhancement", Ben follows medicine in distinguishing between enhancement and repair, drawing into critical view the costs – especially societal costs – due the over-emphasis away from reparative, pro-social applications of technology and towards selfish, personal performance oriented applications. In the competitive business environment, wherein pressures for performance are high, what are the implications of a 'pay *to be able to* play' ethics of human enhancement? One may become a better stockbroker due to an integration with an intelligent machine, or through "nanocognition" for example, but is this making one also a better person? A better world? Opening with review of ethical approaches to these issues, the broad concern for the transformative influence of technology on human beings remains at the fore. Ben's chapter affords a critical reflection on immediate business culture, as concerns for autonomy, sustainability, and virtue are shelved in the face of threats from competitors, yet it is in this environment that everyday people – us - live, adapt or die, and express their moral judgments. Through our collaboration, we make it this way.

Jai Galliot's chapter, our eighth, with welcome candor braves the so-called "responsibility gap" between consequence and locus of responsibility as it appears to be enlarged through the distancing afforded intelligent machines in an already cloudy field of automated warfare. Jai writes that moral responsibility is located in agents which "intentionally make a free and informed causal contribution to any act in question, meaning that they must be aware of the relevant facts and consequences of their actions, having arrived at the decision to act independently of coercion and were able to take alternative actions based on their knowledge of the facts" and moves from here to discuss ways in which technologically mediated agency including semi-autonomous war machines complicate matters. For example, he deftly isolates the ways in which distant drone operators are unable to form "a view of the 'bigger picture' … perhaps limiting responsibility" and rather pursues again a central theme in preceding contributions, that we integrate the machines as we integrate others with limited responsibility, as partly responsible. Jai writes that "We need to move away from the largely insufficient notion of individual responsibility, upon which we typically rely, and move towards a more complex notion of collective responsibility, which has the means and scope to include non-human action." One way to smooth this integration is to level morality down to a common denominator, and Jai endorses such a pass "primarily" because "classical accounts raise endless questions concerning free will and intentionality" insoluble "from a practical perspective aimed at achieving results here and now." His approach is instead to "conceive of moral responsibility as less of an individual duty and more of a role that is actively defined by pragmatic group norms" with the upshot, on his account, being that we can then begin to ascribe responsibility to institutions and machine agents for the roles that they fulfill in achieving consequences worthy of moral reprobation, and with "the greatest share of responsibility … ascribed to the most capable agents."

Where Jai's focus is on military applications of intelligent machines and the ethical distance afforded human operators thereof, the next and ninth chapter takes on the ethical implications of distancing, and de-distancing, technologies in the field of education. The situation as described by authors James Willis III and Viktoria Strunk is clear, with discourse in "ethics in technology, specifically at the granular level of learning analytics … at an intellectual crossroads." And, their warning for educators especially is no less dire: "Unless principled ideas are brought within the public sphere of technological development, the speed of scientific innovation will render ethics of technology misguided at best and obsolete at worst." For ethical discourse to become practically mute within the field of education is troubling as technology affects contemporary public and for-profit education in increasingly suspect ways. For instance, authors Willis and Strunk point out that "in near live-time, administrators, faculty members, and researchers are able to assess a student's activity, engagement, and potential outcome, through predictive algorithms". Bentham's "panopticon" comes immediately to mind, as this technology affords the potential to dramatically reduce the costs of oversight around few distanced administrators. This trend is further exacerbated by MOOCs and AI, whereby the monetization of education encourages the removal of human educators from the financial equation altogether. Willis and Strunk note:

In late 2014, a company that provides online professional development notified its adjunct instructors that their current positions will cease to exist due to computer-mediated, algorithmic response to discussion posts, supplemented with "peer-to-peer" dialogue. While the tenets of computer-based interaction and peer-to-peer assessment may be debated, the replacement of human expertise with computer-generated responses gives personal evidence of the replacement of the scholar. The age of automated teaching is nigh.

Running through this chapter's discussion is the sense of education – and life – as an art, as opposed to the sense of education as a sort of machining, and of course this wing of the division runs through Dewey and Peirce to the Greeks. From this traditional stance, Willis and Strunk propose that the goal of education for educators remains for "students to learn from their mistakes and be able to apply those skills." However, with education coming increasingly under the influence of increasingly automated economies of scale, pressures from "the now accepted *business* model of education", alongside social pressures to deliver education in relatively easily quantifiable pursuits like Science, Technology, Engineering and Mathematics (STEM), are mounting against the retention of this professional, practical, and ultimately democratic political attitude. Willis and Strunk write: "Because there is money to be made in education, it is easily accepted for many in the business world to think of students as customers, education as a service/end product, and faculty as replaceable entertainers." This monetization of education has infiltrated scholarship, as professors are hired to teach but are evaluated on (quantity of) publications. One casualty has been the quality of published scholarship, with scandals involving falsified experimental data common enough. Instead of rewarding pro-social education, the monetization of education has turned "publish-or-perish" into "publish-and-profit", a model fitting for a novelist, but undeserving of the academy as traditionally conceived. Finally, "pay-to-*open*-publish" digital platforms favor those with the financial resources to afford offering personal journal publications for free across the Internet. These being easily accessed to those without access to adequate university libraries come to dominate the field of ideas along with their authors, and so the model further deteriorates into simple finance. The academy is ruled by demagogues who personally profit from its podiums and the influence over culture that this represents, rather than satisfying the original, democratic aims of education.

The tenth chapter in this collection is a standout, placed here as an extension of the discussion on education from the realm of AI-assisted human education to that of the human assisted education of AIs. Opening with the recognition that research into AMAs is "not only important for equipping [artificial] agents with the capacity of making moral judgments, but also for helping us better understand [human] morality through the creation and testing of computational models of ethical theories", authors Ari Saptawijaya and Luis Pereira build an especially readable ethical bridge from individual morality to collective human ethics out of the aged timber of evolutionary game theory, resulting in a sprawling yet deeply grounded *tour de force* on human and artificial morality. They begin by reviewing foundational work in artificial agency and applications in morally significant contexts (e.g. medicine), and from there move quickly to review their logic programming (LP) approach to modeling morality, with the discussion illustrating how their work captures many hard points present in contemporary ethical discourse. They demonstrate how LP is able to articulate "trolley problems" popular in fields of experimental moral philosophy and neuro-ethics (typically imaging) studies, for example. Their framework, as in the previous chapter with Pereira, chapter 4, is deceptively simple as it quite powerfully captures these inherently dramatic moral decision spaces. This section on trolley problems is especially effective. Saptawijaya and Pereira are able to render the logic driving their models of individual moral reasoners in extremely clear terms while taking advantage of the narratives inherent in everyday discourse over similar moral dilemmas. To this compelling portrait of independent moral reasoning, the second section on collective morality contributes the lesson that "Added dependency on cooperation makes it more competitive to cooperate well" thus making it, for any individual agent, "advantageous to invest in shared morals in order to attract partners who will partake of mutual and balanced advantages." Their review of research in ethics at this level of organization covers some fascinating ground, including for example an evolutionary appropriation of conscience as traditionally understood. That Ari and Luis are able to integrate the individual with the collective consistently with traditional terms like conscience via traditional symbol-pushing LP models is a fascinating discussion. Readers primarily interested in this degree of analytical clarity would be well-advised to begin this volume with this chapter, number 10.

The eleventh chapter is from Rick Searle, co-editor and also author of the first chapter. Rick's entry serves as a capstone for this section on autonomy, education, evolution and violence. "Robots in Warfare and the Occultation of the Existential Nature of Violence" takes on now well-established positions on the ethics of artificial agents in the industry of war. Early on in his discussion, Rick fixes on an insight into the nature of the problem, writing that "the essential ethical question" is missed "by too strong a focus on technology" to such a degree that the "state of the technology decides the ethical questions" for us, as if – once the right tech is in place – we can finally turn our morality over to machines designed to do the work for us.

Rick is especially suspicious of the high-tech business of warfare, noting that a different set of incentives arise for "a military made up of human soldiers where the goal is to finish a conflict as quickly and with the least human damage (at least to one's own side)" than that arises for "manufactures of military robots" incentivized "to generate as much revenue as possible during a conflict" and even accompanied with "a perverse incentive to encourage conflict." Taking the just war tradition as his touchstone, Rick considers contributions of intelligent machine technologies to the processes leading up to and during war, drawing special attention to the democratic responsibility of free human beings to not engage in unjust wars regardless. To sharpen this point, he reminds us of the human costs of war, and not in dead but in terms of what had traditionally been referred to under the heading of "spiritual" or "psychic" damage. War breaks hearts, ruins lives, empties futures. And, war-enabling technologies, they afford a numbing distance from the purposeful ruin. Rick is able to trace the depths of implications to which I point now through current literature, coming to the question - from where arises the moral disgust required to stop systematic violence if not from the bereaved bellies of generations of morally wounded? In this way, Rick brings home one of the deep threads running through the contributions to this volume, that scholarship at every level eventually grounds out in the world that it helps to shape.

The final chapter of this volume comes from Aleksandar Malecic of the University of Nis, Serbia, and is the most difficult of the collection. The chapter attempts to articulate an elegant application of Aristotle's four causes across apparently different levels, or manifestations, of the material universe including human consciousness. Aleksandar sets out from the notion of a "strange loop" which describes the progress of a system as it moves upwards or downwards in a hierarchy, e.g. levels of organizational complexity, only to arrive at where it left. His discussion winds through an interesting review of quantum theories and gravitation, but strange loops are common to moral fables, old cultural lessons that remain in everyday practice, e.g. "rock, paper, scissors". Thus, this approach comes at the essentially embodied and embedded – situated – character of cognition in a way that preserves figurative connections with mystical and mythical appropriations of the human condition. Moreover, it adds something essential to naturally arising consciousness that challenges similar ascription to an artificial agent. Aleksandar writes:

Consciousness is not an epiphenomenon of Newtonian physics. Metabolism and repair in living beings are necessary ingredients of any really self-aware (containing the model of the environment and itself within the environment) entity capable to adapt its own point of view (context, formal cause) according to known circumstances.

The portrait here given is interesting because it effectively posits that consciousness arises from the embodied, embedded and anticipatory structure of cognition, and that this structure is itself common to all natural systems. Accordingly, Aleksandar writes that "Strange loops aren't a recipe for the creation of conscious machines in a causal world, so much as a requirement."

Aleksandar makes the case that artificial consciousness as typically conceived is impossible. Consciousness arises with purpose, purpose with final cause, and final cause – as essential to the form of any natural consciousness – necessarily involves a relationship between that arising consciousness and its unique end. Aleksandar writes:

In order to have a self-aware computer, one needs to figure out the way to create a program with "strange" causal loops. Since over thirty years after this idea was proposed there still aren't such self-aware algorithms and machines, one has to ask what is wrong with it.

Aleksandar's discussion is complicated by its breadth of implication. For one, it confronts us with the likelihood that our business with technology occludes an enlightening window on our natural condition which, so long as this view is blocked, denies an opportunity to reform our technological projects with this natural condition in view. In aiming for something other than consciousness in a machine, we run the risk of leveling human experience down to those processes obvious in the artificial instantiation thereof. We become intelligent machines, by default, unless – as Aleksandar so energetically proposes – we recognize that in this anticipatory structure of cognition, and agency, we have the power to choose those ends towards which we move. We have the power to direct ourselves, and to coordinate with others in the achievement of ends that are good, if not for the business of technology in its every established application, then for those who are born to live in terms of it.

Jeffrey White
Korean Advanced Institute of Science and Technology, KAIST, South Korea