Foreword

The scientific and technological innovations that are driving the development of expert systems and, perhaps one day, genuine 'artificial intelligence' are rapidly bringing humanity to the brink of civilizational change. This is no idle claim. When combined with other technologies in the fields of robotics, synthetic biology and nano-manufacturing (to name but a few), it is possible to envisage a world in which the vast majority of humanity enjoy access to the material necessities of life without the burden and drudgery of toil. Of course, this 'utopian' vision has been advanced by others before. However, those visions have typically depended on a fictional account of what might be the fruits of human ingenuity. The difference this time, is that those fruits are already being harvested.

Of course, imagined futures also include far darker, 'dystopian' possibilities. There is no need to canvass, here, their general lines. It is, instead, it important to note that the fears that they inspire are already gaining traction. If they occupy a place in the minds of enough people, then this could lead to otherwise useful developments being restrained or abandoned.

So, where does the difference lie in terms of which possible future will emerge?

As I have argued elsewhere:

Technical mastery divorced from ethical restraint is at the root of all tyranny.

We have seen the truth of this maxim expressed in events throughout human history: the weaponization of nuclear physics, the application of superb engineering to create maximally efficient death camps, the list is long.

Too often, the malicious use of science and technology has been enabled by the absence of even the most minimal source of ethical restraint. That is, people have not even realized that there were fundamental ethical issues to be considered. Only after the proverbial 'horse had bolted' – and the damage had been done – would someone pause to ask, "how was this possible? What were we thinking? Oh, the horror!".

It is against this background – of both risk and opportunity – that we should welcome this edition of essays on the general topic of ethics and artificial intelligence (AI). A book of this kind is an invitation not only to reflection but to responsibility. It allows us to pause, if only for a moment, to consider not only if something *can* be done but also whether it *should* be done.

Such a pause is, in itself, potentially revolutionary – especially if the opportunity to reflect is built into the structure of scientific and technological developments in this area. For my part, I think it essential that the topics covered in this book be used to inform the structure of development and application. Indeed, I think that emerging technologies (such as blockchain) should be used to track the ethical 'provenance' of AI systems – from the moment of inception, to design, application and modification.

Each stage should require a statement of the core purpose, values and principles that have been taken into account and most importantly, how they have been given practical effect in the design, development and implementation phases. Any subsequent change should also register any variation from the patter of ethical design. Finally, all of this should be open to assurance by disinterested professionals (tech auditors) who van verify that the stated ethics framework is being expressed in functional terms.

Why Argue for such a regime? It's not simply a response to the growing power of the technology (although that is a factor). Rather, it's more to do with the unusually 'opaque' character of systems like unsupervised neural networks. We can know the initial conditions, we can know if the system works but, in principle, we cannot precisely describe every step in the process (or overlapping processes) that produce a successful outcome. Not knowing the detail necessarily wakens the 'chain of *accountability'*. This the need for a higher degree of *responsibility* on the part of those who create and deploy these technologies – from the outset.

The authors who have contributed to this book have started at 'ground level' launching the trajectory of their shared perspectives from a solid grounding in human rights. From there, the 'sky is the limit' – reaching areas where authonomous expert systems may be directing the affairs of human beings – in effect acting as our 'boss'.

As others have noted, in the past, we should recognise that human beings are no less susceptible to the effects of bias, error, fallible judgement than the machines we are developing. Already, there are some functions where AI significantly outperforms when compared to humans attempting the same tasks. For example, AI exceeds the capacity of all humans in diagnosing, quickly and accurately, some forms of cancer. However, I would argue that only another human can place their hand on another's shoulder to convey the results of a diagnosis – especially if the most likely consequence is otherwise premature death. The difference lies in the fact that only another human being knows what it is to be mortal and therefore can convey, with genuine sympathy, the prospect of death. Yes, an adequately trained system can offer a perfectly convincing performance of such a conversation – but it will always lack knowledge of a truth that makes the world of difference.

There is some debate about whether or not AI will be able to handle genuine ethical dilemmas when, in fact, values and principles (and all that flow from them) are so perfectly balanced as to make the 'right' decision impossible to discern. Humans encounter this all the time – and often manage to resolve such dilemmas by resorting to an option that is a pure act of creation; something so unexpected as to be unpredictable. I mention this final question because it brings us back to the central purpose of this book.

It is a point so obvious as to risk being ignored … this book has been written by humans for humans to read. That is, it begins with a recognition of the fact that we are entirely, inexcusably responsible for the things we make and use. So, let's hope we absorb the impact of that lesson. Let's hope that we think before we act; that we anticipate (as best we can) all that might follow from the choices we make.

As noted at the outset, I believe that AI will transform our world. With books like this at hand, I hope for a better future arising out of a transformation that is both just and orderly. Get the ethics right, and we will be able to look back at this time of decision without regret.

Simon Longstaff
School of Business, University of New South Wales, Australia