The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation

Joanna J. Bryson

July 28, 2019

Abstract

Artificial intelligence (AI) is a technical term often referring to artifacts used to detect contexts for human actions, or sometimes also for machines able to effect actions in response to detected contexts. Our capacity to build such artifacts has been increasing, and with it the impact they have on our society. This does not alter the fundamental roots or motivations of law, regulation, or diplomacy, which rest on persuading humans to behave in a way that provides sustainable security for humans. It does however alter nearly every other aspect of human social behaviour, including making accountability and responsibility potentially easier to trace. This chapter reviews the nature and implications of AI with particular attention to how they impinge on possible applications to and of law.

Keywords: artificial intelligence; regulation; law; diplomacy; accountability; security; social behaviour

For many decades, Artificial Intelligence (AI) has been a schizophrenic field pursuing two different goals: an improved understanding of computer science through the use of the psychological sciences, and an improved understanding of the psychological sciences through the use of computer science. Although apparently orthogonal, these goals have been seen as complementary since progress on one often informs or even progresses the other. Indeed, we have found two factors which have proven to unify the two pursuits. First, the costs of computation and indeed what is actually computable are facts of nature that constrain both natural and artificial intelligence. Second, given the constraints of computability and the costs of computation, greater intelligence relies on the reuse of prior computation. Therefore to
the extent that both natural and artificial intelligence are able to reuse the findings of prior computation, both can be advanced at once.

Neither of the dual pursuits of AI entirely readied researchers for the now glaringly evident ethical importance of the field. Intelligence is a key component of nearly every human social endeavour, and our social endeavours are most of our activities for which we have explicit, conscious awareness. Social endeavours are also the purview of law, and more generally of politics and diplomacy. In short, everything humans deliberately do has been altered by the digital revolution, as well as much of what we do unthinkingly. Often this alteration is in terms of how we can do what we do, for example how we check the spelling of a document, book travel, recall when we last contacted a particular employee, client or, politician, plan our budgets, influence voters from other countries, decide what movie to watch, earn money from performing artistically, discover sexual or life partners, and so on. But what makes the impact ubiquitous is that everything we have done, or chosen not to do, is at least in theory knowable. This fundamentally alters our society because it alters not only what we can do, but how and how well we can know and regulate ourselves and each other.

A great deal has been written about AI ethics recently. Much of it unfortunately has not focussed either on the science of what is computable, nor on the social science of how ready access to more information and more (but mechanical) computational power has altered human lives and behaviour. Rather, a great deal of it has focussed on AI as a thought experiment or ‘intuition pump’ through which we can better understand the human condition or the nature of ethical obligation. This volume is focussed on the law—the day-to-day means by which we regulate our societies and defend our liberties. This chapter sets context for the volume by introducing AI as an applied discipline of science and engineering.

**Intelligence is an ordinary process**

For the purpose of this introduction I will use an exceedingly-well established definition of intelligence, dating to the seminal monograph on animal behaviour. Intelligence is the capacity to do the right thing at the right time. It is the ability to respond to the opportunities and challenges presented by a context. This simple definition is important because it demystifies intelligence, and through it AI. It clarifies both intelligence’s limits and our own social responsibilities, in two ways.

---

First, note that intelligence is something that operates at a place and in a moment. It is a special case of computation, which is the physical transformation of information\(^1\). Information is not an abstraction\(^2\). It is physically manifested in light, sound, or materials. Computation and intelligence are therefore also not abstractions. They require time, space, and energy. This is why—when you get down to it—no one is really ever that smart. It is physically impossible to think of everything. We can make tradeoffs: we can for example double the number of computers we use and cut the time of a computation nearly in half. Not quite in half, because there is always an extra cost\(^3\) of splitting the task and recombining the outcomes of the processing. But this requires double the space for our two computers, and double the energy in the moment of computation, though the sum of the total energy used is again nearly the same, with the addition of the energy for the overheads. There is no evidence that quantum computing will change this cost equation fundamentally: it should save on not only time but also space, but the energy costs are poorly understood and to date look fiendishly high.

Second, note that the difference between intelligence and artificial intelligence is only a qualifier. Artificial means that something has been made through a human process. This means by default that humans are responsible for it. The artefact actually even more interesting than AI here is a concept: responsible. Other animals can be trained to intentionally limit where they place (for example) even the fairly unintentional byproducts of their digestive process, but as far as we know only humans have, can communicate about, and—crucially—can negotiate an explicit concept of responsibility.

Over time, as we recognise more consequences of our actions, our societies tend to give us both responsibility and accountability for these consequences—credit and blame depending on whether the consequences are positive or negative. AI only changes our responsibility as a special case of changing every other part of our social behaviour. Digital technology provides us with better capacity to perceive and maintain accounts of actions and consequences, so it should be easier not harder to maintain responsibility and enforce the law. However, whether accountability is easier with AI depends whether and in

---


what ways we deploy the capacities digital technology affords. Without care and proper measures, the increased capacity for communication that information communication technology (ICT) provides may be used to diffuse or obscure responsibility. One solution is to recognise in law that the lack of such care and measures for promoting accountability in processes concerning digital artefacts is a form of negligence. Similarly, we could declare unnecessary obfuscation of public or commercial processes a deliberate and culpable evasion of responsibility.

Note that the simplicity of the definitions introduced in this section is extremely important as we move towards law and regulation of systems and societies infused with AI. In order to evade regulation or responsibility, the definition of intelligence is often complicated in manifestos by notions such as sentience, consciousness, intentionality and so forth. I will return to these issues below, but what is essential when considering AI in the context of law is the understanding that no fact of either biology (the study of life) nor computer science (the study of what is computable) names a necessary point at which human responsibility should end. Responsibility is not a fact of nature. Rather, the problem of governance is as always to design our artefacts—including the law itself—in a way that helps us maintain enough social order so that we can sustain human flourishing.

AI—including machine learning—occurs by design

AI only occurs by and with design. AI is only produced intentionally, for a purpose, by one or more members of human society. That act of production requires design decisions concerning at a minimum the information input to and output from the system, and also where and how the computation required to transform that information will be run. These decisions entail also considerations of energy consumption and time that can be taken in producing as good a system as possible. Finally, any such system can and should be defended with levels of both cyber and physical security appropriate to the value of the data transmitted or retained, and the physical capacities of the system if it acts on the world.

The tautology that AI is always generated by design extends to machine learning (ML), which is one means of developing AI wherein computation is used to discover useful regularities in data. Systems can then be built

\[\text{Note that these observations show how under-informed about basic systems engineering the idea of a machine converting the world into paperclips is, as per Nick Bostrom.}\]

to exploit these regularities either to categorize, make predictions, or select actions directly. The mere fact that part of the process of design has been automated does not mean that the system itself is not designed. The choice of ML algorithm, the data fed into it to train it, the point at which it is considered adequately trained to be released, how that point is detected by testing, whether that testing is ongoing if the learning continues during the system’s operation—all of these things are design decisions that not only must be made, but that also can easily be documented. As such, any individual or organisation that produces AI could always be held to account by being asked to produce documentation of these processes.

Despite the fact documentation of such decisions and records of testing outcomes are easy to produce, good practice is not always followed\(^5\). This is as much a matter for the law as any other sloppy or inadequate manufacturing technique\(^6\). What process is deemed adequate for commercial products or even private enjoyment is determined by some combination of expertise and precedent. Whether these processes have been followed and documented can easily be checked either before a product is licensed, after a complaint has been made, or as a part of routine inspection.

Although actual algorithms are abstractions, that only means algorithms in themselves are not AI. In computer science, an algorithm is just a list of instructions to be followed, like a recipe in baking\(^7\). Just as a strand of DNA in itself is not life—it has no capacity to reproduce itself—so instruction sets require not only input (data) but also physical computation. Without significant complex physical infrastructure to execute their instructions, both DNA and AI algorithms are vacuous. The globally-largest technology corporations have almost inconceivably vast infrastructure for every aspect of storing, processing, and transmitting the information that is their business. This infrastructure includes means to generate electric power and provide secure communication as well as means to do computation.

These few leading corporations further provide these capacities also as service infrastructure to a significant percentage of the world’s other ICT companies—of course, at a cost. The European Union (EU) has committed to investing substantial public resources in developing a localised equiva-

---

7. The term algorithm is currently often misused to mean an AI system by those unclear on the distinction between design, program, data, and physical computing system e.g. Cathy O’Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016, p. .

lent of this computational infrastructure resource, as they have previously
done with both commercial aviation and global positioning systems. The
EU may also attempt to build a parallel data resource, though this is more
controversial. There has also been some discussion of ‘nationalising’ signif-
icant technology infrastructure, though that idea is problematic given that
the Internet is transnational. Transnationalising technology ‘giants’ is also
discussed further below.

Digital technology empowers us to do all sorts of things, including ob-
fuscating or simply deleting records or the control systems they refer to. We
can make systems either harder or easier to understand using AI8. These
are design decisions. The extent to which transparency and accountability
should be required in legal products is also a design decision, though here
it is legislators, courts, and regulators that design a regulatory framework.
What is important to realize here is that it is perfectly possible to mandate
that technology be designed to comply with laws, including any ensuring
traceability and accountability of the human actions involved in the design,
running, and maintenance of intelligent systems. In fact, given that the
limits of ‘machine nature’ are far more plastic than those of human nature,
it is far more sensible to minimise the amount of change to laws and rather
maximise the extent of required compliance to and facilitation of extant
laws9.

The performance of designed artefacts is readily ex-
plainable

Perhaps in the desire to evade either laws of nations or the above-mentioned
laws of nature, many deeply respected AI professionals have claimed that
the most promising aspects of AI would be compromised if AI were to be
regulated10. For example, the idea that maintaining standard rights to

8Kroll et al., “Accountable Algorithms”.
9Joanna J. Bryson, Mihailis E. Diamantis, and Thomas D. Grant. “Of, for, and by
the people: the legal lacuna of synthetic persons”. In: Artificial Intelligence and Law
of Robotics. The United Kingdom’s Engineering and Physical Sciences Research Council
themes/%20engineering/activities/principlesofrobotics/.

10My assertion about the ‘deeply respected’ relates to claims I’ve heard in high-level pol-
icy settings, but haven’t been able to find in print. However, for examples of the rhetoric
see Cassie Kozyrkov. “Explainable AI won’t deliver. Here’s why.” In: Hackernoon (Nov.
explanation—demonstration of due process—would eliminate the utilisation
of many advanced machine learning techniques, because these are too com-
plex for their exact workings to be knowable. This last sort of assertion fails
to take into account the present standards for accountability in corporate
law. If a company is audited, that audit never reaches to explaining the
workings of the brain synapses or gene regulation of that company’s em-
ployees. Rather, we look for audit trails—or perhaps witnesses—indicating
that humans have followed appropriate procedures.

AI may reduce the number of people who can be put on a witness stand
to describe their recollections of events or motivations, but it enables a
standard of record keeping that would be unbearably tedious in non-digital
processes. It is not the case that all AI systems are programmed to keep
such records, nor that all such records are maintained indefinitely. But it is
the case that any AI system can be programmed for this, and programmed
using good systems of logging of the design, development, training, testing,
as well as the operation of the systems. Further, individuals or institutions
can choose how, where, and how long to store this logging data. Again, these
are design decisions for both AI systems and the institutions that create
them. There are available standards for adequate logging for generating
proof of due diligence or even explanation of behaviour is. Norms of use for

What matters for human justice is that humans do the right things.
We do not need to check exactly how a machine learning algorithm works
any more than we need to completely understand the physics of torque
to regulate bicycle riding in traffic. Our concerns with AI are that it is
used in a way that is lawful. We want to know for example that products
comply to their claims, that individual users are not spied upon or unfairly
disadvantaged, that foreign agencies were not able to illicitly insert false
information into training a machine learning data set or into a newsfeed.

All AI affords the possibility of maintaining precise accounts of when,
how, by whom, and with what motivation it has been constructed. Indeed,
this is true of artefacts in general, but digital artefacts are particularly
amenable to automating the process. The very tools used to build the sys-

tem can also be set to capture and prompt for this kind of information. We can similarly track the construction, application, and outcomes of any validating tests. Further, even the most obscure AI system after development can be treated entirely as a blackbox and still tested to see what variation in inputs varies the outputs\textsuperscript{12}. Even where performance is stochastic, statistics can tell us the probability of various outcomes, again a sort of information to which the law is already accustomed, for example in medical outcomes.

In fact though almost no AI systems are entirely opaque. Systems with AI are generally far less opaque than human reasoning and less complex than a government or ecosystem. There is a decades-old science of examining complex models by using simpler ones, which is already accelerating to serve the sectors that are already well regulated which are of course (like all sectors) increasingly using AI\textsuperscript{13}. And of course many forms of AI, built either with or without the use of ML, readily produce explanations themselves\textsuperscript{14}.

To return to one of the assertions at the beginning of this section, it is also wrong to assume that AI is not already regulated. All human activity, particularly commercial activity, occurs in the context of some sort of regulatory framework\textsuperscript{15}. The question is how to continue to optimise this framework in light of the changes to society and its capacities introduced by AI and ICT more generally.

\textsuperscript{12}This process is coming to be called (as of this writing) forensic analysis, see e.g. Joseph R. Barr and Joseph Cavanaugh. “Forensics: Assessing model goodness: A machine learning view”. In: Robotic Intelligence. 2019, pp. 17–23. DOI: 10.1142/9789811203480_0003. URL: https://www.worldscientific.com/doi/abs/10.1142/9789811203480_0003.


Intelligence increases by exploiting prior computation

The fact that computation is a physical process limits how much can be done de novo in the instant during which intelligence must be expressed—when action must be taken to save a system from a threat or to empower it through an opportunity. For this reason, much of intelligence exploits computation already done, or rather artefacts produced that preserve the outcomes of that computation. We can understand the outcomes not only of culture but of biology this way. It is not only that organisms can only exploit opportunities they can perceive, it is also that they tend only to perceive what they are equipped to exploit—both capacities for perception and action evolve together. Similarly, culture passes us the tools that others have not only invented but, of all those inventions, the ones that produce the greatest impact relative the costs of transmission, where such costs include both time (suggesting missed opportunities) and the likelihood of adequately faithful replication. Culture itself evolves because that gives us and it efficacy.

Much of the recent immense growth of AI has been due to improved capacities to ‘mine’ using ML the existing discoveries of humanity and nature more generally. The result of this is of course that the good comes with the bad. We mine not only knowledge but stereotypes—and if we allow AI to take action, prejudice—when we mine human culture. This is not


19Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. “Semantics derived automatically from language corpora contain human-like biases”. In: Science 356.6334
a special feature of AI, as mentioned above this is how nature works as well\textsuperscript{20}. Further, show that at least some of what we call ‘stereotype’ reflects aspects of present-day conditions, such as what proportion of jobholders for a particular position have a particular gender. Thus some things we have agreed are bad (e.g. that it is sexist to expect programmers to be male) are aspects of our present culture we have therefore implicitly agreed we wish to change.

One theory for explaining the explosion in what we recognise as AI (that is, of AI with rich, demonstrably human-like, and previously human-specific capacities like speech production and comprehension, or face recognition) is less a consequence of new algorithms than of new troves of data and increased computation speeds. We can expect such explosions of capacities based on the strategy of mining past solutions to soon plateau, when artificial and human intelligence come to be sharing nearly the same, though still-expanding boundary of extant knowledge. In fact, we can also expect this boundary to be expanding faster now, given the extra computational resources we are bringing not only through digital hardware, but by increasing access to other human minds. For humanity, ICT reduces the aforementioned overheads of combining concurrent search. We all get smarter as our culture expands to embrace more—and more diverse—minds\textsuperscript{21}. However, the fact that we can exploit our own computation to build AI, or that we can build our own native as well as systemic intelligence by using AI, does not mean that we are replaceable with AI. As will be explained in the next sections, AI cannot be used to replicate humans, and this has substantial consequences for law and regulation.


AI cannot produce fully replicated humans (all models are wrong)

When computer science is mistaken for a branch of mathematics, many important implications of computation being a physical process are lost. For example, AI is wrongly perceived as a path towards human immortality. First, the potential of ‘uploading’ human intelligence in any meaningful sense is highly dubious. Technologically, brains cannot be ‘scanned’ and replicated in any other material than another brain, as their computational properties depend on trillions of temporal minutiae. Creating a second, identical human to host that new brain is not only physically intractable, but would be cloning—both unethical and illegal, at least in the European Union. Second, even if we could rather somehow upload adequate abstractions of our own minds, we should not confuse this with actually having spawned a digital replica. A digital clone might be of use for example to offload canned email replies, or to create somewhat interactive interfaces for historical storytelling.

Many have argued that the moral intuitions, motivations, even the aesthetics of an enculturated ape can in no way be meaningfully embedded in a device that shares nothing of our embodied physical (‘phenomenological’) experience. Nothing we build from metal and silicon will ever share our phenomenology as much as a rat or cow, and few see cows or rats as viable vessels of our posterity. Yet whether such digital artefacts are viewed as adequate substitutes for a real person depends on what one values about that person. For example, if one is valuing one’s own capacity to control

---


23 as some would suggest, see Murray Shanahan. The technological singularity. MIT Press, 2015, for a review.


the lives of others, many already turn to the simple technology of a will to control sometimes quite intimate aspects of the lives of those chosen to be their heirs. Thus it seems clear that there will be those who spend millions or even billions of dollars, euro, or rubles on producing digital clones they are literally deeply invested in believing to be themselves\textsuperscript{27}.

Even if we could somehow replicate ourselves in an artefact, the mean time to obsolescence of digital technologies and formats is far, far shorter than the average human life expectancy, presently nearing ninety years. This quick obsolescence is true not only of our physical technology, but also of our fashion. Unquestionably any abstracted digital self-portrait would follow fashion in reflecting an aspect of our complex selves that will have been culturally-appropriate only in a specific moment. It would not be possible from such an abstraction to fully model how our own rich individual being would have progressed through time, let alone through biological generations. Such complete modelling opposes the meaning of abstraction. An unabstracted model would again require biological cloning, and even there after many generations would fall out of ecological ‘fashion’ or appropriateness as evolution takes its course.

With apologies to both Eisenhower and, all abstractions are wrong, but producing abstractions is essential. By the definition used in this chapter, intelligent action is an abstraction of the present context, therefore producing an abstraction is the essence of intelligence. But that abstraction is only a snapshot of the organism, it is not the organism itself.

Reproducing our full organism is not required for many aspects of what calls ‘positive immortality’. Replicating our full selves is certainly not essential to writing fiction or otherwise making a lasting contribution to a culture or society, or an irrevocable impact on the ecosystem. But the purpose of this chapter is to introduce AI from the perspective of maintaining social order — that is, from the perspective of law and regulation. As will be discussed below, the methods for enforcing law and regulation are founded on the evolved priorities of social animals. Therefore any intelligent artefacts representing such highly abstracted versions of an individual human

\textsuperscript{27}Pasquale, “Two concepts of immortality: Reframing public debate on stem-cell research”, questions such expenditures, or even those of in vitro fertilisation, on the grounds economic fairness.  
\textsuperscript{27}Pasquale, “Two concepts of immortality: Reframing public debate on stem-cell research".
are not relevant to the law except perhaps as the intellectual property of their creator.

**AI itself cannot be dissuaded by law or treaty**

There is no way to ensure that an artefact\(^{28}\) could be held legally accountable. Many people think the purpose of the law is to compensate, and obviously if we allow a machine to own property or at least wealth then it could in some sense compensate for its errors or misfortune. However, the law is really primarily designed to maintain social order by dissuading people from doing wrong. Law dissuades by making it clear first what actions are considered wrong, and second what are the costs and penalties for committing these wrong acts. This is even more true of policy and treaties, which are often constructed after long periods of negotiated agreement between peers or at least sufficiently powerful actors about what these wrongs and costs are. The Iran Nuclear Deal\(^{29}\) is an excellent example of this.

Of course all of these systems of governance can also generate revenue, which may be used by governments to some extent to right wrongs. However, none of the costs or penalties that courts can impose will matter to an AI system. While we can easily write a program that says “Don’t put me in jail!” we cannot program the full, systemic aversion to the loss of social status and years of one’s short life that the vast majority of humans experience by birthright. In fact, not only humans but many social species find isolation and confinement deeply aversive—guppies can die of fright if separated from their school, and factory farming has been shown to drive pigs to exhibit symptoms of severe mental illness\(^{30}\).

We might add a bomb, camera, and timer to a robot, and program the bomb to destruct if the camera has seen no humans (or other robots) for ten minutes. Reasoning by empathy you might think this machine is far more dissuaded than a human, who can easily spend more than ten minutes alone. But empathy is a terrible system for establishing universal ethics—it


works best on those most like you\textsuperscript{31}. The robot’s behaviour could easily be utterly unaltered by this contrivance, and so it could not be said to suffer at all by the technical definitions of suffering\textsuperscript{32}. Even if the robot could detect and reason about the consequences of its new situation, it would not feel fear, panic, or any other systemic aversion, although depending on its goals it may alter its planning to favor shorter planning horizons.

Law has been invented by—we might even say ‘coevolved with’—our societies to hold humans accountable, thus only humans can be held accountable with it. Even the extension of legal personality to corporations only works to the extent that real humans who have real control over those corporations suffer if the corporation does wrong. The overextension of legal personhood to corporations that are designed to fail is called making a ‘shell company’. Similarly, if you build an AI system and allow it to operate autonomously, it is essential that the person who chooses to allow the system to operate autonomously is the one who will go to jail, be fined, etc. if the AI system transgresses the law. There is no way to make the AI system itself accountable. AI being itself held accountable would be the ultimate shell company\textsuperscript{33}.

The implicit principles that underly our capacity to coordinate and cooperate through the law and its dissuasions have also coevolved with our advanced societies. We share many of our cognitive attributes—including perception and action capacities, and importantly, motivations—with other apes. Yet we also have specialist motivations and capacities reflecting our highly social nature\textsuperscript{34}. No amount of intelligence in itself necessitates social competitiveness, neither does it demand the desire to be accepted by an ingroup, to dominate an outgroup, nor to achieve recognition within an ingroup. These are motivations that underlie human cooperation and

\textsuperscript{31}Paul Bloom. \textit{Against empathy: The case for rational compassion}. Random House, 2017.


\textsuperscript{33}Bryson, Diamantis, and Grant, “Of, for, and by the people: the legal lacuna of synthetic persons”.

competition that result from our evolutionary history\textsuperscript{35}. None of this is necessary—and much of it is even incoherent—from the perspective of an artefact. Artefacts are definitionally designed by human intent, not directly by evolution. With these intentional acts of authored human creation\textsuperscript{36} comes not only human responsibility, but an entirely different landscape of potential rewards and design constraints\textsuperscript{37}.

AI and ICT impact every human endeavour

Given that AI can always be built to be explainable, and that only humans can be held to account, assertions that AI itself should be trustworthy, accountable, or responsible are completely misguided. If only humans can be held to account; from a legal perspective the goal for AI transparency is to ensure that human blame can be correctly apportioned. Of course there are also other sorts of transparency such as those that support ordinary users in establishing the correct boundaries they have with their systems (defending their own interests), and for developers or other practitioners to be able to debug or customize an AI system\textsuperscript{38}. AI can be reliable but not trustworthy—it should not require a social compact or leap of faith\textsuperscript{39}.

Consumers and governments alike should have confidence that there is a fact

\begin{itemize}
\item \textsuperscript{36} The choice to create life through childbirth is not the same. While we may author some of childrearing, the dispositions just discussed are shared with other primates, and are not options left to parents or other conspecifics to determine.
\item \textsuperscript{38} Bryson and Winfield, “Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems”.
\end{itemize}
of the matter that they can determine at will about who is responsible for the AI infused systems we incorporate into our homes, our business processes, and our security.

Every task we apply our conscious minds to—and a great deal we do implicitly—we do using our intelligence. Artificial Intelligence therefore can affect everything we are aware of doing and a great deal we have always done without intent. As mentioned earlier, using even fairly trivial and ubiquitous AI we recently demonstrated that human language contains our implicit biases, and further those in many cases reflect our lived realities. In reusing and reframing our previous computation, AI allows us to see truths we hadn’t previously known about ourselves, including how we transmit stereotypes, but it doesn’t automatically or magically improve on us without effort. Caliskan, Bryson, and Narayanan show also that the outcome of the famous study showing that given otherwise-identical resumés, individuals with stereotypically African American names were half as likely to be invited to a job interview as individuals with European American names. Smart corporations are now using carefully-programmed AI to avoid the implicit biases at the early stages of their HR processes to get the sort of diverse CVs to the short-list stage where hiring decisions can—with explicit care and intention—avoid perpetuating the mistakes of our past.

The idea of having ‘autonomous’ AI systems ‘value aligned’ is therefore also likely also misguided. While it is certainly necessary to acknowledge and understand the extent to which implicit values and expectations must be embedded in any artefact, designing for such embedding is not sufficient to create a system that is autonomously moral. Indeed, if as I have argued a system cannot be accountable, it may also not in itself be held as a moral agent. The issue should not be ‘embedding’ our intended (or asserted) values in our machines, but rather ensuring that our machines can

40Caliskan, Bryson, and Narayanan, “Semantics derived automatically from language corpora contain human-like biases”.
41Lewis and Lupyan, “Language use shapes cultural norms: Large scale evidence from gender”.
allow the expression of the mutable intentions of ourselves, their operators.

Only through correctly expressing our intentions should AI incidentally telegraph our values. Individual liberty, including freedom of opinion and of thought, are absolutely critical not only to human well being but to a robust and creative society\(^43\). Allowing ‘values’ to be enforced by the enfold-
ing curtains of interconnected technology invites gross excesses by powerful actors against those they consider vulnerable, a threat, or just unimpor-
tant\(^44\). Even supposing a power that is demonstrably benign, allowing it the mechanisms for technological autocracy creates a niche that may facilitate a less-benign power— whether through change of hands, corruption of the original power, or just corruption of the systems communicating its will. Finally, who or what is a powerful actor is also altered by ICT, where clandestine networks can assemble—or be assembled\(^45\)—out of small numbers of anonymous individuals acting in a well coordinated way, even across borders.

Theoretical biology tells us that where there is greater communication, there is a higher probability of cooperation\(^46\). Cooperation has nearly entirely positive connotations, but it is in many senses almost neutral—nearly all human endeavours involve cooperation, and while these generally benefit many humans, some are destructive to many others. Further, the essence of cooperation is moving some portion of autonomy from the individual to a group\(^47\). The extent of autonomy an entity has is the extent to which it determines its own actions\(^48\). Individual and group autonomy must to some


\(^{45}\)Carole Cadwalladr. “I made Steve Bannon’s psychological warfare tool’: meet the data war whistleblower”. In: *The Observer* (18 March 2018).


\(^{47}\)Bryson, “Artificial Intelligence and Pro-Social Behaviour”.

extent trade off, though there are means of organising groups that offer more or less liberty for their constituent parts.

Many people are (falsely) preaching that ML is the new AI, and (again falsely) that the more data ML is trained on, the smarter the AI. ML is actually a statistical process we use for programming some aspects of AI. Thinking that bigger data is necessarily better begs the question, better for what? Basic statistics teaches us that the number of data points we need to make a prediction is limited by the amount of variation in that data, providing only that the data is a true random sample of its population\textsuperscript{49}. So there are natural limits for any particular task on how much data is needed—except perhaps for surveillance. What we need for science or medicine may require only a minuscule fraction of a population. However, if we want to spot specific individuals to be controlled, dissuaded, or even promoted, then of course we want to “know all the things.”\textsuperscript{50}

The changing costs and benefits of investment at the group level that Roughgarden, Oishi, and Akçay describe has other consequences beyond privacy and liberty. ICT facilitates blurring the distinction between customer and corporation, or even the definition of an economic transaction. Customers now do real labour for the corporations to whom we give our custom: pricing and bagging groceries, punching data at ATMs for banks, filling in forms for airlines and so forth\textsuperscript{51}. The value of this labour is not directly remunerated—we assume that we receive cheaper products in return, and as such our loss of agency to these corporations might be seen as a form of bartering. ‘Free’ services like search and email may better be understood as information bartering\textsuperscript{52}. These transactions are not denominated with a price, which means that ICT facilitates a black or at least opaque market reducing both measured custom and therefore tax revenue. This is true for everyone who uses Internet services and interfaces, even ignoring the present controversies over definitions of employment raised by platforms\textsuperscript{53}. Our in-

\textsuperscript{49}Meng18.


\textsuperscript{51}Bryson, “Artificial Intelligence and Pro-Social Behaviour”.


\textsuperscript{53}though see Tim O’Reilly. \textit{WTF? What’s the Future and why It’s Up to Us}. New York:
ability to assign value to these transactions may also explain the mystery of why AI doesn’t seem to be increasing productivity.\(^5^4\)

AI then gives us new ways to do everything we do intentionally and a great deal else. The extent to which AI makes different tasks easier and harder varies in ways that are not intuitive. This also increases and decreases the values of human skills, knowledge, social networks, personality traits, and even locations, and alter the calculations of identity and security. Fortunately, AI also gives us tools for reasoning and communicating about all these changes, and for adjusting to them. But this makes group-level identity itself more fluid, complicating our ability to govern.

**Who’s in charge? AI and governance**

Despite all of this fluctuation, there are certain things that are invariant to the extent of computational resource and communicative capacities. The basic nature of humans as animals of a certain size and metabolic cost, and the basic drives that determine what gives us pleasure, pain, stress, and engagement are not altered much. How we live is and always will be enormously impacted by how our neighbours live, as we share geographically-related decisions concerning investment in air, water, education, health, and security. For this reason there will always be some kind of geography-based governance. The fundamental ethical framework we have been negotiating globally for the last century or so of human rights is based on the responsibility of such geographically-defined governments to individuals within the sphere of influence of that government.\(^5^5\) Now wise actors like the European Union have extended the notion of their individual’s sovereignty over cyber assets such as personal data.\(^5^6\) This makes sense for almost exactly the same reason as rights to airspace make sense. With bidirectional information access we can influence individual’s behaviour just as we could with physical force.

---


Recently there has been good reason to hope that we really will start mandating developers to follow best practice in software engineering. If we are sensible, we will also ensure that the information systems spreading and engulfing us will also be entirely cybersecure (or else not on the Internet), with clearly-documented accountability and lines of responsibility. Nevertheless, even if these visions can be achieved, there are still other areas of law and governance with which we should be concerned. The last one I focus on in this present chapter are the new centres of power and wealth. As just explained in the previous section, these are also parts of the everything human that AI and ICT are altering. Further, it is clear that achieving secure and accountable AI requires cooperation with adequate sources of power to counter those who wish to avoid the consensus of the law. Therefore wealth and power distribution, while again like cybersecurity clearly orthogonal technologically to AI, are also clearly irrevocably intertwined with its ethical and regulated application. Problems of AI accountability and grotesquely uneven wealth distribution are unlikely to be solved independently.

In this section it should be noted that I am describing my own work in progress with colleagues, but some of it seems sufficiently evident to justify inclusion now. For example, we hypothesise that when new technologies reduce the economic cost of distance, this in turn reduces the amount of easily-sustained competition in a sector. This is because locale becomes less a part of value, so higher quality products and services can dominate ever-larger regions, up to and including the entire globe. Such a process may have sparked the gross inequality of the late Nineteenth and Early Twentieth Centuries, when rail, news and telecommunication, and oil (far easier to transport than coal or wood) were the new monopolies. Inequality spirals if capital is allowed to capture regulation, as seems recently to have happened not only with ‘big tech’ globally, but also for example with finance in the UK or oil in Saudi Arabia and Russia, leading to a ‘resource curse.’

---


60 John Christensen, Nick Shaxson, and Duncan Wigan. “The Finance Curse: Britain
the mid-Twentieth Century, stability was only reclaimed via the innovation of the welfare state, which in some countries (including the US and UK) preceded at least the second World War, though cooperation sadly there too was motivated by the first.

Governance can be almost defined by redistribution; certainly allocation of resources to solve communal problems and create public goods is governance’s core characteristic. Thus excessive inequality can be seen as a failure of governance. Right now what we are clearly not able to govern (interestingly, on both sides of the Great Firewall of China) are Internet companies. As a result, and similar to the market for commercial aircraft, the costs of distance are sufficiently negligible that the best products are very likely to become global monopolies, unless there is a substantial government investment, e.g., the Great Firewall of China, or Airbus in Europe. Where governance fails in a local region e.g. a county is also where we are likely to see political polarisation and populist candidates or referendum outcomes.

Many problems we associate with the present moment then were not necessarily created by AI or ICT directly, but rather indirectly by increasing inequality and regulatory capture. Other problems were not so much created and the World Economy”. In: The British Journal of Politics and International Relations 18.1 (2016), pp. 255–269. DOI: 10.1177/1369148115612793. URL: https://doi.org/10.1177/1369148115612793; Nolan M McCarty, Keith T Poole, and Howard Rosenthal. Polarized America: The dance of ideology and unequal riches. second. Cambridge, MA: MIT Press, 2016.


E.g. a Gini coefficient over 0.27 Francesco Grigoli and Adrian Robles. Inequality Overhang. IMF Working Paper WP/17/76. International Monetary Fund, 2017, note that too low can be problematic too.


as exposed by AI\textsuperscript{66}. There are some exceptions where ICT—particularly, the capacity of digital media to be fully reproduced at distance inexpensively—do produce qualitative change. These include changing of the meaning of ownership\textsuperscript{67}, and generating truly novel means for recognising and disrupting human intentions, even implicit ones not known by their actors\textsuperscript{68}. On the other hand, some things are or should be treated as invariant. As an example mentioned earlier, human rights are the painstakingly agreed foundation of international law and the obligations of a state, and should be treated as core to ethical AI systems\textsuperscript{69}.

One of the disturbing things we come to understand as we learn about algorithms is the extent to which humans are ourselves algorithmic. Law can make us more so, particularly when we constrain ourselves with it, for example with mandatory sentencing. But ordinarily, humans do have wiggle room\textsuperscript{70}. As mentioned earlier, trust is based on ignorance\textsuperscript{71}—that ignorance may be an important feature of society that ICT might remove. Trust allows cheating or innovating, and sometimes this may be essential. First, allowing innovation makes the level of detail about exceptions that needs to be specified more tractable. Second, of course, innovation allows us to adjust to the unexpected and to find novel, sometimes better solutions. Some—perhaps many—nations may be in danger of allowing the digital era to make innovation or free thinking too difficult or individually risky, creating nation-wide fragility to security threats as well as impinging on the important human


\textsuperscript{70}Cohen, “What privacy is for”.

right of freedom of opinion\footnote{Cf. Frischmann and Selinger, \textit{Re-engineering humanity}, p. }. In such countries, law may bend too much towards the group and inadequately defend the individual. As I mentioned, this is an issue not only of rights, but also of robustness—individuals and variation produce alternatives, and choosing amongst these is a rapid way to change behaviour when a crisis demonstrates change is needed\footnote{Cohen, “What privacy is for”; Luke Stark. “The emotional context of information privacy”. In: \textit{The Information Society} 32.1 (2016), pp. 14–27. doi: 10.1080/01972243.2015.1107167. \url{https://doi.org/10.1080/01972243.2015.1107167}.}. Given that the digital revolution has fundamentally changed the nature of privacy for everyone, all societies will need to find a way to reintroduce and defend ‘wiggle room’ for innovation and opinion. I believe strongly that it would be preferable if this is done not by destroying access to history, but by acknowledging and defending individual differences, including shortcomings and the necessity of learning. But psychological and political realities remain to be explored and understood, and may vary by polity.

\section*{Summary, and the robots themselves}

To reiterate my main points, when computer science is mistaken for a branch of mathematics, many important implications of computation being a physical process are lost. Further, the impact on society of the dissemination of information, power, and influence has not noted been adequately noted in either of those two disciplines, while in law and social sciences, awareness of technological reality and affordances has been building only slowly. Ironically, these impacts until very recently been noticed much in political science. Primarily, these impacts were most noted only in sociology, which was unfortunately in many ways imploding at the same time AI was exploding. Similarly to the myopia of computer science, psychology has primarily seen itself as studying humans as organisms, and the primary ethical considerations in that field were seen as being those of medical subjects, such as patient privacy. Again, some related disciplines such as media studies or marketing raised the issue that as we better understood human behavior we might more effectively manipulate and control it, but that observation made little headway in the popular academic understanding of artificial intelligence. Direct interventions via neuroscience and drugs received more attention, but the potential for indirect manipulations particularly of adults were seemingly dismissed.

These historic errors may be a consequence of the fact that human adults
are of necessity the ultimate moral agents. We are the centres of accountability in our own societies, and as such are expected to have the capacity to take care of ourselves. AI ethics therefore was often reduced to its popular culture edifice as an extension of the civil rights movement. Now that we have discovered—astonishingly!—that people of other ethnicities and genders are as human as ‘we’ are, ‘we’ are therefore obliged by some to consider that anything might be human. This position seems more a rejection of the inclusivity of civil and human rights than an appropriate extension, but it is powerfully attractive to many who seem particularly likely to be members of the recently dominant gender and ethnicity, and who perhaps intuit that such an extension would again raise the power of their on clique by making the notion of rights less meaningful.

More comprehensibly, some have suggested we must extend human rights protections to anything that humans might identify with in order to protect that self concept, even if it is implicit or mistaken. This follows from Kant’s observation that those that treat animals reminiscent of humanity badly are also more likely to treat humans badly. Extending this principle to AI is also most likely a mistake, and an avoidable one. Remember that AI is definitionally an artefact and therefore designed. It almost certainly makes more sense where tractable to change AI than to radically change the law. The UK first and now very recently the OECD have recommended that AI should not deceptively appear to be human, so such confusions might be minimised. This may seem heavily restrictive at present, but as society becomes more familiar with AI—and through that process, better understands what it is about being human that requires and deserves protection—we should be able to broaden the scope of how near to human

---

76 Boden et al., Principles of Robotics.
77 OECD, Recommendation of the Council on Artificial Intelligence.
devices can be while still not having them be deceptive\textsuperscript{78}.

As discussed earlier, there are recent calls to ground AI not on ‘ethics’ (which is viewed as ill defined) but on international human rights law. Of course, this may be a false dichotomy; procedures from classical ethics theories may still be of use in determining ambiguities and tradeoffs of application\textsuperscript{79}. We can certainly expect ongoing consideration of localised variation which ethics perhaps better encapsulates than rights. Ethics has always been about codes of conduct, which confound fundamental principles which we may be able to codify with rights, with other things that are essentially identity markers. But identity too can be essential to security through constructing a defendable identity\textsuperscript{80}. Identity obviously (definitionally) defines a group, and groups are often the best means humans have for achieving security and therefore viability. Not only is breaking into different groups sometimes more efficient for governance or other resource constraints, but also some groups will have different fundamental security tradeoffs based on their geological and ecological situation, and also just simply their neighbours. That identity often also rests on shared historical narratives which will afford different organisational strategies as well may be secondary to the more essential geo-ecological concerns (as is illustrated by the apparent ease with which new ethnicities are invented\textsuperscript{81}) it still of course also makes a contribution.

In conclusion, any artefact that transforms perception to more relevant information, including action, is AI—and note that AI is an adjective, not a noun, unless it is referring to the academic discipline. There is no question that AI and digital technologies more generally are introducing enormous transformations to society. AI Nevertheless, these impacts should be governable by less transformative legislative change. The vast majority of AI—


particularly where it has social impact is and will remain a consequence of corporate commercial processes, and as such subject to existing regulations and regulating strategies. We may require more regulatory bodies with expertise in examining the accounts of software development, but it is critical to remember that what we are holding accountable is not machines themselves, but the people who build, own, or operate them—including any who alter their operation through assault on their cybersecurity. What we need to govern is the human application of technology; what we need to oversee are human processes of development, testing, operation, and monitoring.

AI also offered us an opportunity to discover more about how we ourselves and our societies work. By allowing us to construct artefacts that mimic aspects of Nature but with new affordances for modularity and decoupling, we allow ourselves novel means of self examination, including of our most crucial capacities such as morality and political behaviour. This is an exciting time for scientific and artistic exploration as well as for commerce and law. But better knowledge also offers an opportunity for better control. The role of the law for crafting both individual and societal protections has never been more crucial.

Acknowledgements

A small proportion of the material in this review was derived from a document previously delivered to the OECD (Karine Perset) in May 2017 under the title “Current and Potential Impacts of Artificial Intelligence and Autonomous Systems on Society,” which contributed to the OECD AI policy efforts and documents of 2018–2019, and also reused (with permission) and expanded for. More debt is probably owed to Frank Pasquale for extremely useful feedback and suggestions on a first draft. Thanks also to Will Lowe, Patrick Slavenburg, and Jean-Paul Skeete. I was supported in part by an AXA Research Fellowship in AI Ethics while writing this chapter.

References


Bryson, “The Past Decade and Future of AI’s Impact on Society”.

81

Bryson, Joanna J. “Embodiment versus Memetics”. In: Mind & Society 7.1 (June 2008), pp. 77–94.


Cadwalladr, Carole. “I made Steve Bannon’s psychological warfare tool: meet the data war whistleblower”. In: The Observer (18 March 2018).


Claxton, Guy. Intelligence in the flesh: Why your mind needs your body much more than it thinks. Yale University Press, 2015.


— From Bacteria to Bach and Back. Allen Lane, 2017.


“Can and should robots have rights?” In: *Ethics and Information Technology* 20.2 (2018), pp. 87–99.


McSweeney, Bill. *Security, identity and interests: a sociology of international relations*. 


