DAVID L. POOLE & ALAN K. MACKWORTH

ARTIFICIAL INTELLIGENCE 3E

FOUNDATIONS OF COMPUTATIONAL AGENTS

Chapter 18

The Social Impact of Artificial Intelligence

> *Never in the history of humanity have we allowed a machine to autonomously decide who should live and who should die, in a fraction of a second, without real-time supervision. We are going to cross that bridge any time now, and it will not happen in a distant theatre of military operations; it will happen in that most mundane aspect of our lives, everyday transportation. Before we allow our cars to make ethical decisions, we need to have a global conversation to express our preferences to the companies that will design moral algorithms, and to the policymakers that will regulate them.*
>
> – Awad et al. [2018, p. 63]

Artificial intelligence is a transformational set of ideas, algorithms, and tools. AI systems are now increasingly deployed at scale in the real world [Littman et al., 2021; Zhang et al., 2022a]. They have significant impact across almost all forms of human activity, including the economic, social, psychological, healthcare, legal, political, government, scientific, technological, manufacturing, military, media, educational, artistic, transportation, agricultural, environmental, and philosophical spheres. Those impacts can be beneficial but they may also be harmful. Ethical and, possibly, regulatory concerns, as raised by Awad et al. [2018], apply to all the spheres of AI application, not just to self-driving cars.

Autonomous agents perceive, decide, and act on their own. They, along with **semi-autonomous agents**, represent a radical, qualitative change in technology and in our image of technology. Such agents can take unanticipated actions beyond human control. As with any disruptive technology, there may be substantial beneficial and harmful consequences – many that are difficult to evaluate and many that humans simply cannot, or will not, foresee.

Consider social media platforms, which rely on AI algorithms, such as deep learning and probabilistic models, trained on huge datasets generated by users. These platforms allow people to connect, communicate, and form social groups across the planet in healthy ways. However, the platforms typically optimize a user's

feed to maximize engagement, thereby increasing advertising revenue. Maximizing engagement often leads to adversarial behavior and polarization. The platforms can also be manipulated to drive divisive political debates adversely affecting democratic elections, and produce other harmful outcomes such as **deep fakes**. They can also be very invasive of users' privacy. Automated decision systems, possibly biased, are used to qualify people for loans, mortgages, and insurance policies, and even to screen potential employees.

People expect to have the right to receive fair and equitable treatment, to appeal decisions, to ask for accountability and trustworthiness, and to expect privacy. Is it possible to ensure that those rights are indeed available?

DAVID L. POOLE & ALAN K. MACKWORTH

ARTIFICIAL INTELLIGENCE 3E

FOUNDATIONS OF COMPUTATIONAL AGENTS

18.1 The Digital Economy

The science, the technology and the applications of AI have developed rapidly in the era of ubiquitous digital communication and the Internet. The world economy has been transformed by these developments. Most of the ten largest global corporations, measured by market capitalization, rely heavily upon AI applications. Those companies are centred more on the use of information than on the production of material goods. The shift from matter to information is characterized as **dematerialization**.

Physical mail has been disrupted by email, texting, and social media. Incidentally, email was overwhelmed by spam until AI methods were used to filter it out. Printed books are now supplemented by e-books. Analog photography, film, and video are supplanted by digital media. Some travel in planes and cars has been replaced by digital communication. CDs have been replaced by streaming audio and newspapers by news websites. This process, the **atoms-to-bits** transformation, allows transactions with less friction and more speed. It is easier, quicker, cheaper, and more material and energy efficient to stream music than to go to a store to buy a CD.

Digitalization, in turn, leads to a general temporal speedup of society and the economy. It also shrinks distances through telecommunication. We all live now in a global village, as characterized by Marshall McLuhan [1962]. Furthermore, the digital revolution reduces or eliminates the need for intermediaries, such as retail clerks, bank tellers, and travel agents, between the producers and consumers of goods and services – a process known as **disintermediation**.

The digital revolution and AI are transforming the world economy. These effects are beneficial for some but harmful for others. The benefits, and the harms, are far from evenly distributed in the new economy. There is a **winner-take-all** dynamic as the most powerful corporations use their power to increase their dominance until a monopoly, or oligopoly, market position is established. AI and machine learning algorithms, relying on tracking and modeling users, are central to this dynamic. Zuboff [2019] characterized the new economy as **surveillance capitalism**, epitomized by the large-scale harvesting of personal data online to facilitate targeted monitoring and advertising for commercial and political purposes. Human values such as privacy, dignity, equity, diversity, and inclusion are compromised.

Human attention, selective concentration on available information, is a critical and limited resource. Attention is a psychological issue but it is also an economic issue, as pointed out long ago by Simon [1971] when he created the key concept of the **attention economy**. He observed:

> *In an information-rich world, the wealth of information means a dearth of something else: a scarcity of whatever it is that information consumes. What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a*

poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it.

In English, a person is said to be "paying attention" to a salient event. In other words, attention is a currency to be spent, like money, in this economy. The central role of human attention in our screen-filled digital age is described by Richtel [2014]. Turning attention into a commodity requires monitoring users, which, in turn, triggers privacy concerns. Corporations, and other actors, not only want to know a lot about us but they also use that knowledge to manipulate our attention, our thoughts, and our actions.

18 The Social Impact of Artificial Intelligence ReferencesIndex 18.2 Values and Bias

DAVID L. POOLE & ALAN K. MACKWORTH

ARTIFICIAL INTELLIGENCE 3E

FOUNDATIONS OF COMPUTATIONAL AGENTS

18.2 Values and Bias

Learning systems, trained on large datasets, produce outputs that reflect any bias present in the training sets. Since the datasets were acquired in the past, using them to predict outcomes in the future propagates any bias from the past to the future. What if the future will not, or should not, resemble the past?

In machine learning, **bias** has a neutral technical meaning, "the tendency to prefer one hypothesis over another". The **no-free-lunch theorem** implies that any effective learning algorithm *must* have a bias in that sense. But in ordinary language use, human **bias** has a negative connotation, meaning "prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair" [Stevenson and Lindberg, 2010].

Training sets for **facial recognition**, often acquired without informed consent, typically do not represent people equitably, thereby causing misclassification, often with harmful effect as discussed in Section 7.7. **Large language models**, pre-trained on vast text corpora, when prompted often produce new text that is racist, sexist, or otherwise demeaning of human dignity.

Any AI-based decision system inherently reflects certain implicit values, or preferences. The key question to ask is: whose values are they? Typically, the values embedded in an AI system are the values of the designer or owner of the system, or the values implicit in a deep learning training set. Further questions arise. Can those values be made explicit? Can they be specified? Is it possible to ensure those are democratic values, avoiding discrimination and prejudice? Can they be transparent to the users, or targets, of the system? Do they reflect the values of everyone who may be affected, directly or indirectly? Can systems be designed that respect privacy, dignity, equity, diversity, and inclusion?

The role of social bias in training data is described in Section 7.7. Bias and other social impact concerns in modern deep learning systems trained on large corpora are discussed in Section 8.7. These questions, ongoing challenges to AI system designers, are examined critically by O'Neil [2016], Eubanks [2018], Noble [2018], Broussard [2018], Benjamin [2019], and Bender et al. [2021].

DAVID L. POOLE & ALAN K. MACKWORTH

ARTIFICIAL INTELLIGENCE 3E

FOUNDATIONS OF COMPUTATIONAL AGENTS

18.3 Human-Centred Artificial Intelligence

Mary Wollstonecraft Shelley's *Frankenstein; or, The Modern Prometheus* [Shelley, 1818], is the first true science fiction novel. It can be read as a morality tale, as signaled by Shelley's alternate title, *The Modern Prometheus*. According to ancient Greek mythology, Prometheus stole fire from the gods and gave it to humanity. Zeus punished that theft, of technology and knowledge, by sentencing Prometheus to eternal torment. Dr. Frankenstein's creature attempted to assimilate into human society by learning human customs and language, but humankind rejected him and misinterpreted his genuine acts of kindness. That rejection and his loneliness, given his lack of a companion, led to his choice to exact revenge. Frankenstein's monster has now come to symbolize unbridled, uncontrolled technology turning against humans.

Concerns about the control of technology are now increasingly urgent as AI transforms our world. Discussions of so-called **artificial general intelligence** (**AGI**) envisage systems that outperform humans on a wide range of tasks, unlike so-called "narrow" AI that develops and trains systems for specific tasks. Some believe that AGI may lead to a **singularity** when AGI bootstraps to a **superintelligence**, that could dominate humans [Good, 1965]. Or, as Bostrom [2014] hypothesized, an imagined AGI system, given a goal that includes maximizing the number of paperclips in the universe, could consume every resource available to it, including those required by humans. This seemingly absurd thought experiment purports to show that an apparently innocuous AGI, without **common sense**, could pose an existential threat to humanity if its goals are misspecified or otherwise not aligned with the long-term survival of humans and the natural environment. This **safety** concern has come to be known as the **alignment problem** [Christian, 2020].

A more immediate threat is that AI systems, such as self-driving cars and lethal autonomous weapons, may make life-or-death decisions without meaningful human oversight. Less dramatically, AI systems may make harmful, even if not life-threatening, value-laden decisions impinging on human welfare, such as deciding who should get a mortgage or a job offer. This has given rise to a focus on autonomy and human control. How can designers create **human-centred AI** or **human-compatible AI**? Can human values be instilled in AI systems? These questions are examined by Russell [2019], Marcus and Davis [2019], and Shneiderman [2022]. One proposed technique for incorporating human values is **Reinforcement Learning from Human Feedback** (**RLHF**) [Knox and Stone, 2009]. RLHF is the framework for a key module of **ChatGPT** [OpenAI, 2022].

Increasingly, especially in high-stakes applications, human decision-makers are assisted by **semi-autonomous agents**; this combination is known as **human-in-the-loop**. As shown in Chapter 2, intelligent systems are often structured as a hierarchy of controllers, with the lower levels operating very quickly, on short time horizons, while the higher levels have longer time horizons, operating slowly on more symbolic data. Human interaction with hierarchically structured systems typically occurs at the higher levels. Human drivers cannot meaningfully modify the anti-lock braking systems on a car in real time, but they can provide high-level navigation preferences or directions. Humans can steer or brake to avoid accidents but only if they are paying attention; however, as vehicles become more automated the driver may well be distracted, or asleep, and unable to redirect their attention in time.

The concept of **attention** in neural networks is inspired by the concept of **human attention**. Concepts directly related to human attention include **vigilance**, the state of keeping careful watch for possible danger, and **salience**, the quality of being particularly noticeable or important. Designing AI systems so that humans can meaningfully interact with them requires designers who understand the economic, social, psychological, and ethical roles of vigilance, salience, and attention. Early research on human attention and vigilance is reported by N. H. Mackworth [1948] and J. F. Mackworth [1970]. Mole [2010] presents a philosophical theory of attention. The papers collected in Archer [2022] show how issues concerning salience, attention, and ethics intersect.

Designers of interactive AI systems must be well versed in the principles and practices of both **human-computer interaction** (**HCI**) [Rogers et al., 2023] and AI. Good designs for AI can go a long way in creating trustworthy systems. For example, the "Guidelines for Human-AI Interaction" by Amershi et al. [2019] give strategies for doing less when the system is uncertain to reduce the costs and consequences of incorrect predictions.

Assistive technology for disabled and aging populations is being pioneered by many researchers and companies. **Assisted cognition**, including memory prompts, is one application. **Assisted perception** and **assisted action**, in the form of smart wheelchairs, companions for older people, and nurses' assistants in long-term care facilities, are beneficial technologies. Assistive technology systems are described by Pollack [2005], Liu et al. [2006], and Yang and Mackworth [2007]. **Semi-autonomous** smart wheelchairs are discussed by Mihailidis et al. [2007] and Viswanathan et al. [2011]. However, Sharkey [2008] and Shneiderman [2022] warn of some dangers of relying upon robotic assistants as companions for the elderly and the very young. As with autonomous vehicles, researchers must ask cogent questions about the development and use of their creations. Researchers and developers of assistive technology, and other AI applications, should be aware of the dictum of the disability rights movement presented by Charlton [1998], "Nothing about us without us."

A plethora of concepts are used to evaluate AI systems from a human perspective, including **transparency**, **interpretability**, **explainability**, **fairness**, **safety**, **accountability**, and **trustworthiness**. They are useful concepts but they have multiple, overlapping, shifting, and contested meanings. **Transparency** typically refers to the complete ecosystem surrounding an AI application, including the description of the training data, the testing and certification of the application, and user privacy concerns. But transparency is also used to describe an AI system whose outcomes can be interpreted or explained, where humans can understand the models used and the reasons behind a particular decision. Black-box AI systems, based, say, on deep learning, are not transparent in that sense. Systems that have some understanding of how the world works, using causal models, may be better able to provide explanations. See, for example, this presentation on explainable human-AI interaction from a planning perspective by Sreedharan et al. [2022]. Enhancements in explainability may make an application more trustworthy, as Russell [2019] suggests.

Enhanced transparency, interpretability, and fairness may also improve trustworthiness. Interpretability is useful for developers to evaluate, debug and mitigate issues. However, the evidence that it is always useful for end-users is less convincing. Understanding the reasons behind predictions and actions is the subject of **explainable AI**. It might seem obvious that it is better if a system can explain its conclusion. However, having a system that can explain an incorrect conclusion, particularly if the explanation is approximate, might do more harm than good. Bansal et al. [2021] show that "Explanations increased the chance that humans will accept the AI's recommendation, regardless of its correctness."

As discussed in Section 7.7, models built by computational systems are open to probing in ways that humans are not. Probing and testing cannot cover all rare events, or **corner cases**, for real-world domains. Verification of systems, proving that their behaviors must always satisfy a formal specification that includes explicit **safety** and **goal constraints** could make them more trusted [Mackworth and Zhang, 2003]. **Semi-autonomous** systems that interact and collaborate with humans on an ongoing basis can become more

trusted, if they prove to be reliable; however, that trust may prove to be misplaced for corner cases. The role of explicit utilities in open and accountable group decision making is described in Section 12.6. In Section 13.10, concerns about real-world deployment of reinforcement learning are outlined. Trust has to be earned.

18.2 Values and Bias ReferencesIndex 18.4 Work and Automation

DAVID L. POOLE & ALAN K. MACKWORTH

ARTIFICIAL INTELLIGENCE 3E

FOUNDATIONS OF COMPUTATIONAL AGENTS

18.4 Work and Automation

The impact of automation, in general, and AI, in particular, on the nature of work and employment has been widely studied by economists and sociologists; however, there is no consensus yet on the impact of automation or AI. Some claim that AI, including robotics, will lead to large-scale unemployment; others claim that many new job categories will develop based on AI-enabled developments. A better way to look at the phenomenon is to understand that any particular job requires a suite of skills. Some of those skills may indeed be rendered redundant. Other skills may become more necessary and may also require upgrading. As discussed above, **disintermediation** eliminates many job categories but also requires "upskilling" other job categories. These issues are explored by Brynjolfsson and McAfee [2014], Agrawal et al. [2019], and Ford [2021]. One theme, developed by Agrawal et al. [2022], is that business decisions require prediction and judgment. Machine learning is now enabling automated prediction so human judgment and decision analysis skills become relatively more valuable. Danaher [2021] considers the ethics of automation and the future of work.

AI and related technologies are creating many new job categories; however, the new post-industrial high-tech corporations typically employ many fewer people than corporations based in the older industrial economy, with similar market size. AI is now permeating the entire economy, with AI-related jobs being created in the older industrial corporations, such as the auto industry and other manufacturing sectors as well as in the health, legal, education, entertainment, and financial sectors. One aspect of the role of AI in the video game industry is described in Section 6.6. Perhaps fewer people will be required to produce society's goods and services. Moreover, AI could generate so many significant new wealth opportunities that a **universal basic income** (UBI) guaranteed to everyone, without qualification, is possible, and necessary, to redistribute some of that wealth equitably [Ford, 2021]. The argument for UBI is that AI will reduce the need for much manual and mental labour, so the human rights to housing and sustenance should not be tied entirely to employment income. This could allow more creative leisure time and informal caregiving.

It is already the case that the employment picture is changing significantly, disrupted by AI. Many workers now have a portfolio of employment gigs, working on short-term ad hoc contracts. The so-called **gig economy** allows AI-enabled scheduling and organizing of the resources needed for just-in-time ordering and delivery of consumer goods and services, including ride-hailing and food delivery. This has produced radical changes in the nature of retail shopping and employment. A permanent full-time job with a single employer for life is no longer the standard model. The gig economy has the benefit of flexibility, for both the employee and the employer. On the downside, workers are losing the advantages and protections of organizing in unions, including security of employment, bargaining for wages and salaries, and benefits such as vacations, paid sick leave, pensions and health care coverage (if it is not universal). Enhancements to government legislation, regulation, and enforcement are being proposed to cope with these emerging challenges.

DAVID L. POOLE & ALAN K. MACKWORTH

ARTIFICIAL INTELLIGENCE 3E

FOUNDATIONS OF COMPUTATIONAL AGENTS

18.5 Transportation

Transportation, of people and cargo, is a key sector of the economy, satisfying a variety of social needs. It serves as a useful case study to examine the social and economic impact of AI. **Autonomous vehicles** are being developed and deployed. The technologies used for accurate positioning in self-driving vehicles are covered in Section 9.8. Some of the ethical choices surrounding self-driving cars are considered in Section 2.4. The role of preferences in automated route planning is discussed in Section 3.9. Using constraints to schedule deliveries by a fleet of vehicles is described in Section 4.9. The positive impact of having intelligent cars and trucks could be large [Thrun, 2006]. There is the **safety** aspect of reducing the annual carnage on the roads; it is estimated that 1.2 million people are killed, and more than 50 million are injured, in traffic accidents each year worldwide [Peden et al., 2004]. Vehicles could communicate and negotiate at intersections. Besides the consequent reduction in accidents, there could be up to three times the traffic throughput [Dresner and Stone, 2008].

The improvements in road usage efficiency come both from smarter intersection management and from platooning effects, whereby automated, communicating vehicles can safely follow each other closely because they can communicate their intentions before acting and they react much quicker than human drivers. This increase in road utilization has potential positive side-effects. It not only decreases the capital and maintenance cost of highways, but has potential ecological savings of using highways so much more efficiently instead of paving over farmland, forests, or wilderness.

With full autonomy, elderly and disabled people would be able to get around on their own, without a driver. People could dispatch vehicles to the parking warehouse autonomously and then recall them later. Individual car ownership could become mostly obsolete, when an autonomous taxi ride becomes cheaper and more convenient than a private vehicle. Most private vehicles are used only about 5% of the time. Better utilization of the vehicle fleet would significantly reduce the demand for vehicle production and storage. Supported by AI systems, people could simply order up the most suitable available vehicle for their trips. Automated robotic warehouses could store vehicles more efficiently than using surface land for parking. In very dense cities, private car ownership is already becoming obsolete. This trend would accelerate with autonomous vehicles. Much of the current paved space in urban areas could be used for housing, or for environmentally enhancing uses such as parks, playgrounds, or urban farms. The rigid distinction between private vehicles and public transit could dissolve.

These speculations are, at the moment, mostly science fiction. Many early promises of full autonomy have not materialized. The transition to a mixed transportation system of human drivers, autonomous vehicles, transit, pedestrians, cyclists, and so on is challenging.

Short of full vehicle autonomy, many smart driving features such as self-parking, lane keeping, lane changing, adaptive cruise control, emergency braking, and automated avoidance of pedestrians and cyclists are now routine driver aides and safety enhancements. A variety of vehicles, other than cars and trucks, including microcars, e-bikes, e-scooters, and e-unicycles, are now available under the rubric **micromobility**,

often with AI enhancements for semi-autonomy, routing, and vehicle sharing. Public transit, with intelligent crew and vehicle scheduling, and some autonomy is also improving.

Experimental autonomous vehicles are seen by many as precursors to robot tanks, military cargo movers, and automated warfare. Although there may be, in some sense, significant benefits to robotic warfare, there are also very real dangers. In the past, these were only the nightmares of science fiction. Now, as automated warfare becomes a reality, those dangers have to be confronted. Sharkey [2008], Singer [2009a, b], Russell [2019], and Robillard [2021] discuss the dangers and ethics of autonomous weapon systems and robotic warfare.

18.4 Work and Automation ReferencesIndex 18.6 Sustainability

DAVID L. POOLE & ALAN K. MACKWORTH

ARTIFICIAL INTELLIGENCE 3E

FOUNDATIONS OF COMPUTATIONAL AGENTS

18.6 Sustainability

The sustainability crisis now facing humanity has many facets, including the climate emergency, global inequity, biodiversity loss, and scarcity of water and food resources. **Sustainability** is the ability to maintain the balance of a process in a system over the long term. **Ecological sustainability** is the ability of an ecosystem to maintain ecological processes, functions, biodiversity, and productivity into the future. Ecosystem **resilience** is the capacity of an ecosystem to tolerate disturbance without collapsing into a qualitatively different state. In social systems, resilience is enhanced by the capacity of humans to anticipate and plan for the future [Holling, 1973].

Sustainable development is the ability to recognize and meet the needs of the present without compromising the ability of future generations to meet their own needs. In the United Nations Brundtland Report, "Our Common Future" [Brundtland et al., 1987], sustainable development was emphasized. Sustainable development requires satisfying environmental, societal, and economic **constraints** [Rockström et al., 2009; United Nations, 2015b]. Environmental, social, and economic issues are intertwined.

In *An Essay on the Principle of Population*, Malthus [1798] was concerned primarily with the imbalance between population growth, which has grown exponentially, and the supply of food, which is limited. He wrote:

> *This natural inequality of the two powers, of population, and of production of the earth, and that great law of our nature which must constantly keep their effects equal, form the great difficulty.*

In other words, the global planetary system must satisfy the constraint that the consumption by the growing population is limited by the food production of the Earth. It is just one of many constraints that must be satisfied for our planetary system to be sustainable and resilient.

What is the relationship between sustainability and computation, in general, and AI, in particular? Computation is a double-edged sword with respect to sustainability. The amazing increase in the power of our computational and communication networks has been significantly beneficial to sustainability as the digital age unfolds. Computation is transforming society and the economy. As discussed in Section 18.1, computation has, at its core, an inherent sustainable dynamic, **dematerialization**, replacing atoms by bits. Dematerialization, inherently, saves many resources.

On the other hand, many resources are consumed and wasted in the digital revolution. Mining to produce the materials needed to manufacture computers, devices, and batteries can have serious environmental

effects. At the end of the short product lifecycles, many million tonnes of electronic waste are produced each year, with devastating environmental consequences, especially in the Global South. The power used by massive cloud servers is another major resource consumed. In particular, the training of large models, discussed in Section 8.5.5, requires huge computational resources. AI is characterized as a "technology of extraction" by Crawford [2021]. Similarly, the mining of some **cryptocurrency** coins, such as Bitcoin, and the verification of cryptocurrency transactions are also major resource sinks.

Countering these trends is the so-called **green information technology** movement, which aims to design, manufacture, use, repair, and dispose of computers, servers, and other devices with minimal energy use and impact on the environment.

A new discipline, **computational sustainability**, is emerging [Gomes et al., 2019]. It applies techniques from AI, computer science, information science, operations research, applied mathematics, and statistics for balancing environmental, societal, and economic needs for sustainable development. Computational sustainability has two main themes:

- Developing computational models and methods for **offline** decision making for the management and allocation of ecosystem resources.
- Developing computational modules embedded directly in **online** real-time ecosystem monitoring, management, and control.

AI plays a key role in both themes.

In *Planetary Boundaries: Exploring the Safe Operating Space for Humanity*, Rockström et al. [2009] identified nine critical boundaries on the Earth's biophysical processes to ensure the sustainability of the planet. The boundaries are **goal constraints** on:

- climate change
- rate of biodiversity loss (terrestrial and marine)
- interference with the nitrogen and phosphorus cycles
- stratospheric ozone depletion
- ocean acidification
- global freshwater use
- change in land use
- chemical pollution
- atmospheric aerosol loading.

For example, a constraint on anthropogenic **climate change** requires atmospheric carbon dioxide concentration to be less than 350 ppmv (parts per million by volume). The pre-industrial value was 280 ppmv; in 2009 it was 387 ppmv and 412 ppmv in 2023. The rate of biodiversity loss is determined by the extinction rate (number of species lost per million per year). Its boundary value is set at 10, whereas it is greater than 100 in 2023. **Constraint satisfaction**, as covered in Chapter 4 and Section 6.4, is at the core of computational sustainability.

In 2015, the United Nations adopted the "2030 Agenda for Sustainable Development" [United Nations, 2015a] which specifies 17 **Sustainable Development Goals** (**SDGs**) [United Nations, 2015b]. The SDGs cover the nine biophysical planetary boundary constraints and extend them to cover human social and economic goals such as reducing poverty, hunger, and inequality, while improving health, education, and access to justice. Many systems, using the full spectrum of AI methods, including deep learning, reinforcement learning, constraint satisfaction, planning, vision, robotics, and language understanding, are being developed to help achieve the SDGs. For example, as described earlier, Perrault et al. [2020] show how multiagent techniques based on **Stackelberg security games** can enhance public health, security, and social justice. Multiagent methods also address the so-called **tragedy of the commons**, which is at the heart of sustainability concerns [Hardin, 1968]. Ostrom [1990] showed that institutions for collective action can evolve to govern the commons.

AI researchers and development engineers have some of the skills required to address aspects of concerns about global warming, poverty, food production, arms control, health, education, the aging population, and demographic issues. They will have to work with domain experts, and be able to convince domain experts that the AI solutions are not just new snake oil. As a simple example, open access to tools for learning about AI, such as this book and **AIspace** [Knoll et al., 2008], empowers people to understand and try AI techniques on their own problems, rather than relying upon opaque black-box commercial systems. Games

and competitions based upon AI systems can be very effective learning, teaching, and research environments, as shown by the success of **RoboCup** for robot soccer [Visser and Burkhard, 2007]. Some of the positive environmental impacts of intelligent vehicles and smart traffic control were discussed in Section 18.5. Bakker et al. [2020] present an overview of digital technology applications for dynamic environmental management.

Environmental decision making often requires choosing a set of components that work together as parts of a complex system. A **combinatorial auction** is an auction in which agents bid on packages, consisting of combinations of discrete items [Shoham and Leyton-Brown, 2008]. Determining the winner is difficult because preferences are usually not additive, but items are typically complements or substitutes. Work on combinatorial auctions, already applied to spectrum allocation (allocation of radio frequencies to companies for television or cell phones) [Leyton-Brown et al., 2017], logistics (planning for transporting goods), and supply chain configuration [Sandholm, 2007], could further be applied to support carbon markets, to optimize energy supply and demand, and to mitigate climate change. There is much work on smart energy controllers using distributed sensors and actuators which improve energy use in buildings.

18.5 Transportation ReferencesIndex 18.7 Ethics

DAVID L. POOLE & ALAN K. MACKWORTH

ARTIFICIAL INTELLIGENCE 3E

FOUNDATIONS OF COMPUTATIONAL AGENTS

18.7 Ethics

Moral and **ethical** issues abound in considering the impacts of AI. Morals are guidelines that apply to an individual's sense of right and wrong. Ethical principles apply at the level of a community, an organization, or a profession. Morals and ethics are, of course, intimately connected: an individual may have personal morals that derive from various sources, including the ethical principles of groups they belong to. Normative ethical codes are categorized, by philosophers, as either virtue-based, consequentialist, or deontological [Hursthouse and Pettigrove, 2018]:

- **Virtue** ethics emphasize the values and character traits that a virtuous agent possesses [Vallor, 2021].
- **Consequentialist** (or **utilitarian**) ethics focus on the outcomes of possible actions that the agent can take, measuring the global **utility** of each outcome.
- **Deontological** (or **Kantian**) ethical codes are based on a set of rules the agent should follow.

A focus on **AI ethics** has arisen, motivated, in part, by worries about whether AI systems can be expected to behave properly. Reliance on autonomous intelligent agents raises the question: can we **trust** AI systems? They are not fully trustworthy and reliable, given the way they are built now. So, can they do the right thing? Will they do the right thing? But trust is not just about the system doing the right thing. A human will only see a system as trustworthy if they are confident that it will *reliably* do the right thing. As evidenced by popular movies and books, in our collective unconscious, the fear exists that robots, and other AI systems, are untrustworthy. They may become completely autonomous, with free will, intelligence, and consciousness. They may rebel against us as Frankenstein-like monsters.

Issues of trust raise questions about ethics. If the designers, implementers, deployers, and users of AI systems are following explicit ethical codes, those systems are more likely to be trusted. Moreover, if those systems actually embody explicit ethical codes, they are also more likely to be trusted. The discipline of AI ethics is concerned with answering questions such as:

- Should AI scientists be guided by ethical principles in developing theories, algorithms, and tools?
- What are ethical activities for designers and developers of AI systems?
- For deployers of AI systems, are there applications that should not be considered?
- Should humans be guided by ethical principles when interacting with AI systems?
- Should AI systems be guided by ethical principles, in their interactions with humans, other agents, or the rest of the world?
- What data should be used to train AI systems?
- For each of these concerns, who determines the ethical codes that apply?

AI ethics, as an emerging and evolving discipline, addresses two, distinct but related, issues:

A. **AI ethics for humans**: researchers, designers, developers, deployers, and users.

B. **AI ethics for systems**: software agents and embodied robots.

Each is concerned with developing and examining ethical codes, of one of the three code types, either for humans or for systems.

With regard to AI ethics for humans, many perceive a need for strong professional codes of ethics for AI designers and engineers, just as there are for engineers in all other disciplines. Others disagree. The legal, medical, and computing professions all have explicit **deontological** ethics codes that practitioners are expected or required to follow. For computing, the ACM Committee on Professional Ethics [2018], AAAI [2019], and IEEE [2020] have established ethics codes that apply to their members.

There are several issues around what should be done ethically in designing, building, and deploying AI systems. What ethical issues arise for us, as humans, as we interact with them? Should we give them any rights? There are human rights codes; will there be AI systems rights codes, as well?

Philosophers distinguish among **moral agents**, **moral patients**, and other agents. Moral agents can tell right from wrong and can be held responsible for their actions. A moral patient is an agent who should be treated with moral principles by a moral agent. So, for example, a typical adult human is a moral agent, and a moral patient; a baby is a moral patient but not a moral agent, whereas a (traditional) car is neither. There is an ongoing debate as to whether an AI agent could ever be (or should ever be) a moral agent. Moreover, should current AI systems be considered as moral patients, warranting careful ethical treatment by humans? Presumably not, but is it conceivable that future AI systems, including robots, could be, or should be, ever treated as moral patients? Some of the underlying issues are covered by Bryson [2011], Mackworth [2011], Bryson [2018], Gunkel [2018], and Nyholm [2021]. These issues are partially addressed by the multitude of proposed codes of AI ethics such as those developed by the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems [2019], OECD [2019], and UNESCO [2022].

Facial Recognition

Selinger and Leong [2021], studying the ethics of facial recognition technology, define four forms of **facial recognition**, each with their own risks and benefits:

- **facial detection** finds the location of faces in images, it is common in phones, putting square overlays on faces
- **facial characterization** finds features of individual faces, such as approximate age, emotions (e.g., smiling or sad), and what the person is looking at
- **facial verification** determines whether the person matches a single template; it is used to verify the user of a phone and in airport security
- **facial identification** is used to identify each person in an image from a database of faces; it is used in common photo library software to identify friends and family members.

Facial identification, usually considered the most problematic, has problems that arise both when it is perfect and when it makes mistakes.

If facial identification is perfect and pervasive, people will know they are constantly under surveillance. This means they will be very careful to not do anything that is illegal or anything out of narrow social norms. People's behavior is self-censored, even if they have no intention to commit any wrongdoing. Preventing illegal activity becomes problematic when any criticism of the ruling order or anything that deviates from a narrow definition of normal behavior becomes illegal. Surveillance has a chilling effect that limits self-expression, creativity, and growth, and deprives the marketplace of ideas.

When facial identification makes mistakes, they usually do not affect all groups equally. The error rate is usually much worse for socially disadvantaged people, which can result in those people becoming more targeted.

Given a database of faces, facial identification becomes a combination of facial detection and facial verification. The facial verification on an iPhone uses multiple sensors to build a three-dimensional model of a face based on 30,000 points. It has **privacy-by-design**; the information is stored locally on the phone and not in a server. It has a false-positive rate of 1 in 10 million, which means it is unlikely to have a false positive in normal use. If the same error rate was used on a database of everyone, on average there are about 800

people on Earth who match a particular face. Vision-only techniques have much higher error rates, which would mean that mis-identification would be very common.

People have argued that making facial recognition, in any of its forms, part of normal life provides a slippery slope where it is easy to slip into problematic cases. For example, if a community already has surveillance cameras to detect and prevent crime, it can be very cheap to get extra value out of the cameras by adding on facial recognition.

With regard to AI ethics for systems, how should AI systems make decisions as they develop more autonomy? Consider some interesting, if perhaps naive, proposals put forward by the science fiction novelist Isaac Asimov [1950], one of the earliest thinkers about these issues. Asimov's **Laws of Robotics** are a good basis to start from because, at first glance, they seem logical and succinct.

Asimov's original three laws are:

I. A robot may not harm a human being, or, through inaction, allow a human being to come to harm.
II. A robot must obey the orders given to it by human beings except where such orders would conflict with the First Law.
III. A robot must protect its own existence, as long as such protection does not conflict with the First or Second Laws.

Asimov proposed those prioritized laws should be followed by all robots and, by statute, manufacturers would have to guarantee that. The laws constitute a deontological code of ethics for robots, imposing constraints on acceptable robotic behavior. Asimov's plots arise mainly from the conflict between what the humans intend the robot to do and what it actually does, or between literal and sensible interpretations of the laws, because they are not codified in any formal language. Asimov's fiction explored many hidden implicit contradictions in the laws and their consequences.

There are ongoing discussions of AI ethics for systems, but the discussions often presuppose technical abilities to impose and verify AI system safety requirements that just do not exist yet. Some progress on formal hardware and software verification is described in Section 5.10. Joy [2000] was so concerned about our inability to control the dangers of new technologies that he called, unsuccessfully, for a moratorium on the development of robotics (and AI), nanotechnology, and genetic engineering.

Perhaps the intelligent agent design space and the agent design principles developed in this book could provide a more technically informed framework for the development of social, ethical, and legal codes for intelligent agents.

However, in skeptical opposition, Munn [2022] argues that "AI ethical principles are useless, failing to mitigate the racial, social, and environmental damages of AI technologies in any meaningful sense." But see also "In defense of ethical guidelines" by Lundgren [2023].

18.6 Sustainability ReferencesIndex 18.8 Governance and Regulation

DAVID L. POOLE & ALAN K. MACKWORTH

ARTIFICIAL INTELLIGENCE 3E

FOUNDATIONS OF COMPUTATIONAL AGENTS

18.8 Governance and Regulation

It is increasingly apparent that ethical codes are necessary but not sufficient to address some of the actual and potential harms induced by the widespread use of AI technologies. There are already AI liability and insurance issues. Legislation targeting AI issues is coming into force worldwide. Many countries are now developing AI regulations and laws. A survey of the national AI policies and practices in 50 countries is presented by the Center for AI and Digital Policy [2023]. Issues in robot regulation and robot law are covered in Calo [2014] and Calo et al. [2016].

Zuboff [2019] used the term **surveillance capitalism** to characterize the nexus among AI-based user tracking, social media, and modern commerce. This issue is hard to address solely at the national level since it is a global concern. In 2016, the European Union (EU) adopted the **General Data Protection Regulation** (**GDPR**) [European Commission, 2021] as a regulation in EU law on data protection and privacy, as a part of the human right to privacy regime in the EU. The GDPR has influenced similar legislation in many nations outside the EU. Given the size of the European market, many corporations welcomed the GDPR as giving uniformity to data protection; however, GDPR has not put an end to surveillance capitalism. The EU also adopted the **Digital Services Act** (**DSA**) in 2022 [European Commission, 2022c]. The DSA defines a digital service as any intermediary that connects consumers with content, goods, or other services, including social media. It is designed to protect the rights of children and other users, and to prevent consumer fraud, disinformation, misogyny, and electoral manipulation. There are substantial penalties for infringement.

The OECD AI Principles [OECD, 2019] presented the first global framework for AI policy and governance. In 2022, the EU was debating a draft of the **Artificial Intelligence Act** (**AI Act**) [European Commission, 2022b], the first legislation globally aiming to regulate AI across all sectors. It is designed primarily to address harms caused by the use of AI systems, as explained by Algorithm Watch [2022]. The underlying principle of the AI Act is that the more serious the harms, the more restrictions are placed on the systems. Systems with unacceptable risks are prohibited. High-risk systems must satisfy certain constraints. Low-risk systems are not regulated. For example, social scoring, evaluating individual trustworthiness, would be banned if government-led but not if done by the private sector. Predictive policing would be banned. **Facial recognition** in public places by law enforcement would be restricted. Subsequently, the EU followed up with the AI Liability Directive [European Commission, 2022d, a] which would, if enacted, make it more feasible for people and companies to sue for damages if they have been harmed by an AI system. The US Office of Science and Technology Policy [2022] has developed a "Blueprint for an AI Bill of Rights", a set of five principles and associated practices to help guide the design, use, and deployment of automated systems.

Governance covers government legislation and regulation, **external governance**, but it also refers to **internal governance**, within corporations, government agencies, and other actors who are developing and deploying AI products and services. Many of those actors are putting in place internal governance measures, including ethics codes, to ensure responsible AI guidelines are followed [Amershi et al., 2019; World Economic Forum, 2021]. The cultural and organizational challenges that need to be addressed to create responsible AI systems are described by Rakova et al. [2021]. As a note of caution, Green [2022] suggests, "Rather than

protect against the potential harms of algorithmic decision making in government, human oversight policies provide a false sense of security in adopting algorithms and enable vendors and agencies to shirk accountability for algorithmic harms." Professional standards, product certification, and independent oversight are other means, beyond external and internal governance, to ensure AI **safety**, as discussed by Falco et al. [2021].

The scope of government regulation is hotly debated and subject to intense lobbying efforts. Multinational corporations are alleged to use **ethics washing** to fend off further regulation, arguing that the introduction of internal ethical codes is sufficient to prevent harms. Moreover, the extent of **regulatory capture**, whereby legislators and regulators are influenced by, and aligned with, the corporations they are supposed to regulate, is pervasive. It is a real and significant concern for AI governance.

18.7 Ethics ReferencesIndex 18.9 Review