Chapter 4 in: *R.U.R.* and the Vision of Artificial Life, by Karel Čapek, edited by Jitka Čejková
The MIT Press, 2024. **DOI:** https://doi.org/10.7551/mitpress/14136.001.0001

R.U.R. and the Robot Revolution: Intelligence and Labor, Society and Autonomy

Inman Harvey

God created man in his own image, so some would have us believe. Whatever view one has on that, undeniably a major feature of the contemporary world is the human endeavor to create machines in their own human image. In the context of this volume inspired by Čapek's play *(R.U.R.) Rossum's Universal Robots*, I focus on four aspects of this endeavor.

Any robot must have the ability to affect the world around it, and do so in some sensible nonrandom way. Even if these capacities are very limited, we can dignify them as examples of *intelligence* and *labor*. I class the design and achievement of these capacities to work intelligently as technical issues, matters for engineering solutions. Čapek asserts that his robots have such capacities, but makes no useful observation about them. Most discussion in artificial life and artificial intelligence centers on these engineering problems, but I shall argue that these raise no difficulties in principle. Although there are plenty of fascinating and tough technical challenges that will keep people occupied for centuries, we have at least a rough picture of how to start tackling them.

But Čapek highlights two further aspects of robotics, issues of *society* and *autonomy*, that are more problematic and are not merely engineering issues. A core theme of the play is that the efficiency of robots has a massive economic effect, displacing the livelihoods of human workers. Čapek's subtext was the politics of who benefits from the labor of others (whether human or robot). The autonomy of the labor force (again, whether human or robot) is mediated through social norms and constraints. Your freedom to do what you want lies within the possibilities offered by your social and economic environment—and needs to fit around my freedom to do what I want.

Where do autonomous robots fit into this? In the play, the autonomy of the robots foundered on their inability to reproduce, or recreate further copies, in the absence of skilled human assistance. Autonomy (from the Greek for self-rule, having its own laws) can include an entity being in control of its needed resources (material needs, power resources); in charge of maintaining its own organization (self-repair to counter the ravages of entropy); in charge of its own acts. What *R.U.R.* brings to the fore is the interplay between society and autonomy. At the time it was written, the topic was science fiction. Nowadays the robot revolution is turning into science fact.

Public Perceptions

Public perceptions of artificial life are dominated by two concepts: robots and computers. Both these terms were invented (in their current senses) within the last 100 years, and there are some surprising parallels in their origins.

The Czech word *robota* refers to forced human labor, and this was adapted by Čapek to refer to the mechanical servants of *R.U.R.* So the robots were defined in terms of their subservient economic role alone, their humanlike characteristics going no further than the minimum needed for that role. Somewhat similarly, until the mid twentieth century the word "computer" referred to human office workers, perhaps in a bank or insurance company, occupied on humdrum lengthy and repetitive calculations. When a machine could take over such a role, the term computer carried over from the human worker to the machine. Again, it was only the functional role that mattered—in this case the ability to perform abstract calculations—and any further human characteristics were irrelevant.

Both robots and computers are thus originally defined in terms of severely limited subsets of human capacities, focusing on rather menial work and lacking in wider humanity; but this is sometimes forgotten. Young Rossum's requirements for a robot led him to invent "a worker with the smallest number of needs. . . . He tossed out everything not directly related to the task at hand. . . . They are mechanically far superior to us, they have an astonishing capacity to reason intelligently, but they don't have a soul." A hundred years later some people, unaware of the history of these terms (which were deliberately limited in scope) but puzzled about how human nature relates to these technical artifacts, will perversely ask questions such as "are humans merely robots?" or "merely computers?" This makes no more sense than asking "is butter merely low-fat butter?" What these people probably really want to ask might be better phrased as: “are there any limitations to the range of human capacities that we can build into robots (or computers)?”

The idea of animate machines predates the terms robot and computer by millennia. More than 2,500 years ago Greek mythology tackled artificial life.[1] Homer talked of automata, and later Roman emperors such as Claudius enthused over their commissioned (and very real) automata, powered by pneumatics, used in theatrical spectacles.[2] “Automata” translates from the Greek for something that “acts by itself,” implying something more self-willed than Rossum’s original servant class of robot; when these fictional robots became masters of themselves, achieving self-willed status, it was the transition point at which they became a threat to humans.

Of course, Rossum’s robots were fictional; and the plot has them developing beyond their original limited roles, wanting to become masters rather than servants. In Japan—the country with the most robots—the more usual public perception is that robots are there to assist humans. By contrast, within European and Hollywood culture, the standard trope is for robots to be seen as an enemy threatening to take over humankind. There are at least two predominant versions of this threat.

The first version is that robots will take over all the jobs and leave a dispossessed class of humans; this has elements of truth, was a core concern of the *R.U.R.* play, and is discussed below. The second is that some singularity will be passed, when robot “intelligence” surpasses human “intelligence,” and then the game will be over for humans. This I take to be absurd Hollywood fantasy, based on a confusion between reality and game shows and a naive understanding of what “intelligence” might be. There is no universal interspecies IQ test, and intelligence cannot be summarized in a single number. Rather, it is a somewhat fuzzy term referring to the degree of skill demonstrated by a person (and by extension, an animal or robot) at some set of tasks; as such it is a highly context-sensitive term.

Intelligence

In recent years artificial intelligence (AI) has progressed significantly, with a focus on the intelligence aspect of machines including robots, and specifically on abstract-reasoning intelligence.

[1] Adrienne Mayor, *Gods and Robots: Myths, Machines, and Ancient Dreams of Technology* (Princeton: Princeton University Press, 2018).

[2] Eva Anagnostou-Laoutides and Alan Dorin, “The Silver Triton,” *Nuncius* 33(1) (2018): 1–24.

The straightforward computational side, the systematic performance of routine algorithms that 100 years ago was done by human “computers,” has benefited enormously from hardware developments in speed, in the size of problems that can be tackled, and in the smallness and cheapness of the computing machinery. For some time there has been no contest on straightforward routine computational issues: computers beat humans, hands down.

The areas of intelligence that have presented more of a challenge have been the classes of formal problem solving that have resisted being reduced to algorithms. Many of these relate to pattern recognition, for instance identifying objects in images, or translating spoken Chinese to written English. For decades the orthodox GOFAI approaches—"good old-fashioned AI” that assumed all intelligence was at the core based on some algorithmic machinery equivalent to a Turing machine—made rather disappointing progress. But recent advances based on the competing methodology of neural networks have overtaken GOFAI approaches by leaps and bounds, pursuing a more intuitive concept of intelligence. Rebranded as deep learning—which largely means neural networks at scale, with big data sets—this has driven advances in speech recognition and image analysis that would have seemed incredible a decade ago but are now available to anyone with a smartphone.

Though currently much deep learning technology is implemented on top of a computational substrate, this is not at all necessary, and there may well be a significant shift away from this to different hardware (or wetware) substrates. The methodological move is away from programming a machine with explicit instructions on how to tackle a task, toward setting up the machine so that it learns for itself, from experience, how to tackle the task. AlphaZero, the system that within a few hours of self-play can master (to grand master level) sophisticated games such as chess, shogi, and go, would be the archetypal demonstration of this.[3]

Though a substantial amount of human programming went into jump-starting the ability to learn (and indeed to learn how to learn), by any plausible measure of effort the great majority of the skills developed are attributable to the nonprogrammed learning phase. In the longer term, the jump-start of human programming will fade into irrelevance, a historical detail. Of course, the defenders of GOFAI principles, who want to claim that all cognition is at root reducible to computation, will want to claim that even if a computational jump-start is a relatively small part of deep learning, it is nevertheless crucial and in some sense absolutely required. They are wrong

[3] David Silver, Thomas Hubert, Julian Schrittwieser, et al., “A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go through Self-Play,” *Science* 362(6419) (2018): 1140–1144, DOI: 10.1126/science.aar6404.

when they make such arguments. That is not how natural intelligence in the biological world arose through Darwinian evolution, and evolutionary robotics demonstrates in principle how adaptive intelligent systems can be artificially evolved without the need for any such computational jump-start.[4]

Another argument put forward by the GOFAI advocates, unwilling to accept that they backed the wrong paradigm, is that the skills derived through deep learning are in some sense illegitimate, not real intelligence, because they are opaque. Nobody can explain just how the massively deep AlphaZero network actually ensures such novel and creative wins at go, or chess; it is suggested that this devalues the results. Such an argument fails to acknowledge that exactly the same shortcoming holds for the equivalent human skills. It does not need much reflection to realize that this is inevitable. Our understanding of other people, and of robots and machines, is necessarily limited in scope. Folk psychology gives us practical tools for interpersonal skills, psychology and cognitive science may take us a lot further, but any hope for a complete understanding of how humans work—or indeed of any robots that do a halfway decent job of emulating human performance—is clearly a fantasy. Whatever sense of the word is used, if “X understands X” means no more than “X can do what X can do” it is vacuous; but if it means more, then “X can do more than X can do” is a contradiction.

AI seems to me to have made the transition from being a field of overambitious promises with a poor delivery record to being a normal science that has become so essential to our everyday world that we take it for granted. There will always be the hype merchants, there will always be technical challenges yet to meet; but AI is now safe as an intellectual discipline. The computational GOFAI paradigm, though still fighting a rearguard action, is basically in retreat. We can move on from misguided attempts to see how intelligence can be built up from computations to more rewarding questions such as: How can our ability to compute be built up from our native noncomputational intelligence? How can natural disorganized material, such as “chemical soup,” potentially form a substrate for computational systems?[5]

The methodological foundations of AI are now safe, but the dangers are elsewhere. We do

[4] Inman Harvey, Ezequiel Di Paulo, Rachel Wood, Matt Quinn, and Elio Tuci, “Evolutionary Robotics: A New Science for Studying Cognition,” *Artificial Life* 11(1–2) (2005): 79–98, DOI: 10.1162/1064546053278991.

[5] Matthew D. Egbert, Jean Sébastien Gagnon, and Juan Pérez-Mercader, “From Chemical Soup to Computing Circuit: Transforming a Contiguous Chemical Medium into a Logic Gate Network by Modulating Its External Conditions,” *Journal of the Royal Society Interface* 16(158) (2019), DOI: 10.1098/rsif.2019.0190.

indeed need to face up to the essential and inevitable opaqueness of robots when we consider regulation issues. The same sorts of issues arise in interpersonal regulation, in socialization of ethics and the functioning of legal controls. We return to this below when discussing society and autonomy.

Labor

AI is largely focused on abstract intellectual problems, the kind that humans can do sitting in an armchair. Čapek was focused on robots that did useful labor; no sitting around in armchairs for them. Robots must be material, not abstract; must have physical engagement with the world. This physicality has tended to raise more tricky issues than the intelligence side of the equation.

Notoriously a robot vacuum cleaner is much more likely to be scuppered by the edge of a rug, or dog hairs trapped in the rollers, than it is by any abstract problem of planning a pathway across the room. Practical issues of power requirements and weight are still major problems.

Robotics has been bedeviled by its own version of the GOFAI syndrome, the assumption that any such physical problem can somehow be transformed into a computational problem. But there have been a variety of approaches to embodied robotics that reject this.

Common themes include the realizations that no design can be validated in simulation alone —only tests in the real world count—and that action in the world involves dynamics. Getting robots to walk or ride a bicycle is not a sequence of static issues to be solved but a question of matching the dynamics of forces changing in real time with the effects on the robot-world interfaces; and this has to be reflected in the dynamics of control systems, of "brains." Even though the neural networks of AI are often geared toward static problems, there is a whole field, as yet rather underexploited, of dynamical neural networks.[6]

Robotics has not yet reached the levels of achievement shown in AI: it has been the more backward sibling. The scope and range of possible physical interactions is far wider than for abstract intelligence. Nevertheless, I currently see no obvious barrier in principle to developing robots to achieve any task humans or animals can do. The current apparently insuperable difficulty is with emulating the concomitant ability of biological systems to be self-creating and self-maintaining; we return to this below in the section on autonomy.

[6] Randall D. Beer and John C. Gallagher, "Evolving Dynamical Neural Networks for Adaptive Behavior," *Adaptive Behavior* 1(1) (1992): 91–122, DOI: 10.1177/105971239200100105.

Society

The master-servant relationship is a social one, involving economics and politics. *R.U.R.* is a political play and focuses on these issues more than the scientific ones for robotics. Economic concerns are paramount. The robots are provided with pain—but only so as to protect them against damaging themselves. "Why don't you create a soul for them? . . . That's not in our power. . . . It'd increase production costs."

The efficiency benefits for their human masters are initially viewed optimistically. "In five years the cost of everything will be zero point nothing. . . . There's no more poverty. Yes, the workers will lose their jobs. But by then there will be no more jobs. Everything will be produced by living machines, and humans will only do things they love doing. . . . They'll live only in order to better themselves." Of course, this promise of Paradise does not come to fruition. After a few years "the workers rose up against the Robots and smashed them to pieces . . . various governments created Robot armies." The working people turned out not to share in the economic benefits, and those in power attempted to exploit them so as to maintain their power.

This prescient play from a hundred years ago echoes the sorts of explanations often offered for contemporary political shifts such as Trump's America and Brexit in the UK. Change can leave many dispossessed and disadvantaged; the reactions may be expected to be sometimes ugly. This is the real threat of the robot revolution.

Robert Shiller[7] catalogs numerous historical instances of popular narratives decrying technological developments that threatened to cause unemployment. Ned Ludd was (perhaps) a weaver who in 1779 smashed the knitting frames whose efficiencies had the side effect of putting weavers out of jobs. Whether a real or mythical event, the story led to the nineteenth-century Luddite movement taking his name. Subsequent depressions through to the present have often been attributed in part to labor-saving technology. Clearly the invention of the automobile was going to make redundant a lot of ironmongers making horseshoes; were the new jobs in car mechanics and industry sufficient to compensate?

Shiller (who cites Čapek and *R.U.R.*) cautions that we should acknowledge that such economic narratives may have an influential life of their own, regardless of how factually based

[7] Robert J. Shiller, *Narrative Economics: How Stories Go Viral and Drive Major Economic Events* (Princeton: Princeton University Press, 2019).

they may be. But it is indisputable that automation and technology have already had enormous effects on employment; consider the agricultural revolution that reduces the proportion of people working on the land from greater than 70 to less than 5 percent. In the form of the robot revolution, in both AI and robotics, this can only increase in scope and in speed of change. Unskilled and semiskilled jobs are the first to transition—like those of the human computers who used to work in banking and insurance offices.

Car and truck driving still present challenges, but will eventually be automated. The professions that rely on human judgment (another name for intuition) will fall to advances in AI; radiology and other medical image classification soon, language translation and basic legal practice work soon to follow.

The automation of jobs may have unexpected side effects. For instance, finding a parking space in town is no problem if your car can just cruise around aimlessly while you shop or lunch—but your personal convenience is at what cost in traffic congestion for others? However, it is entirely predictable that any new jobs replacing those taken over will be more skilled rather than less, and in the transition a lot of people will be disadvantaged. Of course, in an ideal world, the optimistic one at the start of *R.U.R.*, the economic benefits of automation could in principle be widely shared; but we know this does not happen inevitably and by magic. The internet, the World Wide Web, promised that when the circulation of information became nearly free, the result would be democratization of such resources. Such good effects have indeed largely happened across the world—but with the unanticipated side effects of corporate control of social media platforms, and the balkanization of political commentary.

Robotics and automation do indeed present very real dangers to mankind. But these are not technological dangers, nor some singularity fantasy of *Terminator* robots seizing control. They are human dangers. If robot missiles are given autonomous powers to fire at will, it is because human governments have so chosen. If economic changes arising from automation result in people being dispossessed, it is because human society has failed to take care of its own. As with climate change, robot change requires political solutions; it depends on who has the power to effect social changes in response to natural events. Both human exploitation of fossil fuels and human ingenuity in fashioning technology are natural events, but we need to face up to the human consequences.

Autonomy

Robots should not need constant monitoring and supervision by humans, and in that sense some degree of autonomy is generally essential to their purpose. How much? A robot (or a person) sent on a space mission to Mars may have no say in what their designated tasks are, yet necessarily must be given some independence, some autonomy, in just how these tasks are to be carried out; after all, it may take tens of minutes to consult by radio with Earth while a response is needed immediately. Some sliding scale of increasing autonomy would seem to go with more independence of choice as to what one chooses to do. But this is within a context of what opportunities and support there are from the surrounding environment.

Beyond the rather extreme example of Robinson Crusoe on a desert island, human autonomy relies upon support from the social and economic systems around one. Without ready access to food in exchange for money, it would fairly quickly become obvious to me just how limited my own autonomy is. Nevertheless, within such interplay between personal freedom and social and environmental constraints there is some scope for autonomy—in principle as much for robots as there is for humans.

Some people think that autonomy is a binary, all-or-nothing property, and that there is some objection in principle to robots having it. The view I take is that it is a relatively innocuous and context-sensitive description that raises no deep philosophical issues. A central heating thermostat is autonomous insofar as it switches the boiler on and off in response to room temperature—but typically not autonomous insofar as somebody else designed it and set it in the room attuned to a specific room temperature. An artificially evolved robot control system[8] is autonomous insofar as it is "responsible for its own actions."

Within the framework of robots interacting with humans, with varying levels of responsibility for their own actions, there is a vast range of possibilities. At one extreme, hitchBOT[9] relied on appealing to the good nature of drivers in hitchhiking 10,000 kilometers across Canada in 2014, keeping a record of its adventures. Similarly reliant on human motive power was Norman White's "Helpless Robot," exhibited at the 1997 European Conference on Artificial Life. Positioned in a gallery, when it senses passing visitors it plaintively appeals to be given a push to rotate one

[8] Harvey et al., "Evolutionary Robotics."

[9] David Harris Smith and Frauke Zeller, "The Death and Lives of hitchBOT: The Design and Implementation of a Hitchhiking Robot," *Leonardo* 50(1) (2017): 77–78, DOI: 10.1162/LEON_a_01354.

way or the other. Once a human-robot dialogue has developed (pushes from the one, requests from the other), the robot becomes increasingly demanding, aggressive even, and then grumpy when abandoned.[10]

These playful examples illustrate that human-robot interaction can take many forms, and even a small degree of autonomy can (literally) travel a long way. The autonomy or freedom to act of a robot (or other human) is only a threat insofar as its exercise diminishes the autonomy of others. This is where social norms and regulations, and their enforcement, become relevant. Undoubtedly these will have to be extended to accommodate robotic advances, with regulation for autonomous driving an obvious immediate concern.

In *R.U.R.*, the area where the Robots' lack of autonomy crucially let them down was their inability—without assistance from knowledgeable and skilled humans—to repair and replace themselves as they degraded and wore out. Presciently, this relates to what may well still be the biggest unsolved issue in artificial life: biological systems are self-creating, self-repairing; can we design material synthetic systems to be the same? We have theoretical approaches, such as autopoiesis, that offer possible routes to understanding what is necessary. In natural evolution there likely was a genuine singularity, when the first autopoietic entity arose that had potential for evolution—and thus instigated the living world we see around us now. But as yet there is no consensus on how to even start to synthesize the equivalent.

Some people will suggest that a similar "we haven't a clue" issue is that of providing a robot with consciousness. Does it make sense to talk of "inflicting pain" on a robot, over and above "inflicting damage," and should this feed through to how we regulate human-robot interaction? One can steal one's neighbor's horse and also inflict suffering on it, two crimes; one can only steal his cabbage. But how about his robot? The consciousness issue is to my mind a philosophical nonissue.[11] It arises from linguistic confusion, from treating subjective consciousness as an objective property that a robot may or may not "have." But "subjective" simply cannot be treated as if objective; that misunderstands the way such concepts work. On the "can robots feel pain" issue I am tempted to follow the same strategy that Turing used on "can machines think": "Nevertheless I

[10] Norman White, "Helpless Robot" installation, on compart: Center of Excellence Digital Art, accessed October 17, 2022, http://dada.compart-bremen.de/item/artwork/609.

[11] Inman Harvey, "Evolving Robot Consciousness: The Easy Problems and the Rest," in *Consciousness Evolving*, ed. Hames H. Fetzer (Amsterdam: John Benjamins, 2002), 205–219.

believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted."[12]

Turing used this in the context of proposing the Turing test, a public comparison between the (anonymized) performance of a computer and a person. His guess as to how far computers would go, by the end of the twentieth century, turned out to be at least roughly in the right ballpark. But his prediction is ultimately phrased in terms of how "the use of words and general educated opinion" will alter. My prediction would be that robot developments for social interaction, both with other robots and with humans, will be aided by visible/audible expressions associated with both physical damage and thwarting of intention, and that taking these as expressions of "pain" and "irritation" will become commonplace and accepted. After some period (Turing's fifty years seems again as good a guess as any) the use of words and general educated opinion will feel comfortable dropping the scare quotes around "pain" and "irritation."

Summary

Those who worry about some terrifying robot apocalypse, triggered by some singularity, will no doubt think the comments above miss the main point. But I consider such fantasies absurd, and as diverting attention from the very real dangers arising from accelerating advances in robotics.

There are fascinating technical challenges in developing the intelligence and physical capabilities of robots, challenges that will no doubt keep people busy for decades and centuries to come. But we know enough to know how to continue tackling them. There are no obvious difficulties in principle now that the computational GOFAI logjam has been convincingly broken.

For decades AI was held back by the confusion between computing as one skill that people could perform, and computing as the mechanism by which this was done. The triumphs of deep learning are not just important as a marker for how we can start to mechanize intuition and learning as well as reasoning, but also as a step toward understanding the relationship between mechanism and performance. This lesson extends well beyond the neural network mechanisms currently used.

What we should be worried about, above all else, is the social, economic, and political impact arising from robotics and automation. The robotic revolution will be at least as widespread

[12] Alan M. Turing, "Computing Machinery and Intelligence," *Mind* 59(236) (1950): 433–460, DOI: 10.1093/mind/LIX.236.433.

as the agricultural and industrial revolutions, and may well happen much faster. Much economic benefit will result, but no doubt this will, if unregulated, be shared unfairly. In consequent social and economic upheavals there will likely be large groups of people made redundant and unable to take full advantage of the new opportunities available; they will be left behind.

Robots taking over human jobs will inevitably have varying degrees of autonomy, and interact with humans in somewhat humanlike fashion. Hence both sides in these interactions will face the normal human issues of when my autonomy interferes with yours. Social norms and regulations will have to be tailored to fit new circumstances. But social norms do not just ease everyday interactions; at a deeper level they also delineate the framework of political power.

A hundred years ago Čapek presciently anticipated this social and economic upheaval in his play *R.U.R.* The coming robot revolution is inevitable. We will need to incorporate the new robot interactions into the social contract(s) of everyday life and, crucially, address the political issues of how economic benefits and power and control are distributed in the aftermath of this revolution.