8. What is 'the Secret of Life'? The Mind-Body Problem in Čapek's *Rossum's Universal Robots (R.U.R.)*

Tom Froese

One of the recurring themes in Čapek's play is the existential question of whether the reductionist materialist worldview – the belief that we can fully explain the world, including ourselves, in terms of nothing but physical processes – can accommodate all that is essential to the human being.

The materialist worldview triumphed with the scientific revolution, which in turn laid the foundations for the military-industrial complex. This historical shift is represented in the play by the business-minded young Rossum inheriting the bio-engineering methodology from the mad scientist old Rossum. A key difference between the two is that old Rossum's materialist stance is an ideological commitment, whereas for young Rossum working within a materialist framework is more a matter of convenience: for him it is sufficient for most practical purposes to replicate the machine-like aspects of a person.

Where does this leave the soul, or what we today might prefer to call consciousness? The question of whether human nature goes beyond its physical aspects, and whether these subjective aspects can also be artificially replicated, is extremely challenging to address in scientific theory and practice –100 years ago as much as now. In addition, it is an open question whether a commercial project that depends on the creation of machine consciousness would even be desirable in the first place. What if those conscious machines turned out to dislike their tasks and spontaneously decided that they will no longer comply with the purpose for which they were designed? Hence, we can understand why, in response to Helena's insistence on the need to incorporate a "soul" into the robots, the scientist Dr Gall responds with "That's not within our power" while the technical director Fabry responds, "That wouldn't be in our interest."

This ambiguity about the reasons for this lack of a complex subjective dimension does not get fully resolved in the story. But no matter whether the robots' shortcoming is intended to reflect an inherent limitation of the materialist approach, or rather a conflict of interest in turning machines into autonomous subjects in their own right, the upshot is that the young Rossum's robots are supposed to be unconscious automata: "There's nothing they're interested in […] They've got no will of their own. No passions. No hopes. No soul.". And yet already at the start of the play we are given clues that the situation may not be so simple:

Hallemeier: […] a couple of times, not very often, mind, they have shown some resistance … […] Or sometimes one of them might suddenly smash whatever's in its hand, or stand still, or grind their teeth – […] It's clearly just some technical disorder.

Domin: Some kind of fault in the production.

Helena: No, no, that's their soul!

Fabry: Do you think that grinding teeth is the beginnings of a soul?

As Helena observes, in these descriptions we potentially see the first stirrings of a subjective point of view. Machines can certainly fail to function, but this does not make them angry. As anyone working with real robots quickly discovers, robots do not care about anything; they just do what they do under whatever conditions – the notion of success or failure only exists in the eyes of external observers. To some extent this shouldn't be surprising, given that neither would you expect a malfunctioning microwave oven to suffer and care about its own predicament. Functional breakdowns of machines may be frustrating to their designers and users, but not for the machines themselves. Helena is therefore on the right track to suggest that a robot which is grinding its teeth may actually be expressing frustration about its situation and, if so, is no longer merely an unconscious machine.

What could be the basis of the robot's incipient capacity to care about things? Čapek provides us with some intriguing clues. First, he chose to envision the robots as composed of an artificial form of biological matter, rather than as standard mechanistic machines. And he already anticipated that it would be practically impossible to design humanoid robots that have all of their knowledge and expertise built-in from the start, and so at least a rudimentary process of learning and apprenticeship is required. Even the need for the development of a rudimentary sense of self seems to be implied:

Domin: … They need to get used to the idea that they exist. There's something on the inside of them that needs to grow or something. And there are lots of new things on the inside that just aren't there until this time. You see, we need to leave a little space for natural development.

The idea that the most promising approach to replicate human intelligence is by allowing the artificial agent to undergo a developmental process like a human child was later famously proposed by the father of the modern computer, Alan Turing, in his classic 1950 article. An open question in this regard is whether the material basis of the artificial agent also makes a difference to its potential for consciousness. The Turing computer was explicitly defined in such a way that it will function identically no matter on which material substrate it is implemented – in a silicon logic machine or a biological neural network. And yet in recent decades there is a growing consensus that Turing's computational formalism is too restrictive to adequately capture all that matters about life and mind. A computer program's substrate independence entails that it exists in a space of pure logic, akin to a Platonic realm of ideas, that is timeless and hence in principle beyond questions of life and death that form the ultimate basis of human existential self-concern.

In contrast to a computer, the very existence of a living being, humans included, is a continual achievement on metabolic and interactive levels whose success cannot be guaranteed in principle. And arguably, it is precisely this irreducible precariousness of the being of a living individual which is why things matter from their perspective at all. In this respect, while Čapek's decision to give his robots a biochemical basis might have seemed misguided in the early computational era of Turing, in the end his choice might turn out to have been rather prescient, especially judging from growing interest in the concrete messiness of organic life in fields such as artificial life, embodied cognition, soft robotics, and synthetic biology.

Could a biochemical substrate make room for a subjective point of view, such as an awareness of frustration and response with anger? We know that the behavior of a Turing machine is completely deterministic and that any apparent "choices" are ultimately forced by the rules specified in its programming. There is simply no operational wriggle room here for free action according to values that could be counterfactually impeded in a way that leads to frustration. It is an interesting open question whether a biological substrate helps us to escape from these logical constraints. It is still certainly the case that if we search for evidence of subjectivity within the body of a living being, such as by dissecting it, we just won't find it; there is just the messy biochemistry. This is clearly expressed in Čapek's play in terms of the notion of "physiological correlates," which comes surprisingly close to the modern scientific concept of the neural correlates of consciousness:

Helena: […] I wanted him to give the robots a soul!

Domin: Helena, it's not a matter of having a soul.

Helena: No, just let me speak. That's what he said as well, he said he could only make physiological changes … alter the physiological …

Hallemeier: The physiological correlates, you mean?

The same point is also made in a more macabre way toward the end of the play when Alquist's gruesome dissection of the robot Damon fails to provide the missing insight into the secret of life. This even more fundamental constraint, namely that looking for the subjectively lived perspective in objectively physical terms will only ever let us measure and manipulate objective correlates rather than the subjective perspective itself, poses a severe challenge to the scientific project of completely explaining, let alone artificially replicating, the full quality of human existence.

Intriguingly, there also seems to be a kind of biological uncertainty principle at play that prevents us from fully knowing even life itself in objective terms. To know the workings of a living body in all its completeness we need to make intrusive interventions in its embodiment, which in the last instance would lead to the death of the individual and leave our knowledge of the living as living incomplete. Thus, paradoxically, while we tend to think of biology as the study of the living, in some respects it is perhaps more appropriately thought of as the study of the no longer living. The "secret of life" is only accessible (if it all) over dead bodies:

DAMON: Do experiments on living robots. Discover how they work!

ALQUIST: Living bodies? You expect me to kill them?

But even if the subjective dimension of consciousness cannot be reduced to objective terms, a fascinating possibility is that it still expresses itself in the gaps of the objective, in those moments when physiological or behavioral events arise that are not completely pre-determined by the system's current material organization or its history of past interactions. In other words, not all events that appear to the external observer of a living body as mere noise may be alike; some unexpected activity may in fact be a marker of subjective involvement in those processes.

From a subjective perspective, this hidden ambiguity of our living embodiment may be experienced as an irresolvable tension at the core of subjectivity. Mind and body are not simply

one, but neither are they exactly two; they are neither one nor two, and hence both incomplete and interdependent. If this is on the right track, then it may turn out that the mind-body gap is not an anomaly of our knowledge – rather, it is a conceptual reflection of the fact that people do not simply coincide with their material body. Thus, the mind-body gap becomes the root of our irreducible freedom as human beings.

But, importantly, this is a kind of freedom that is also irreducibly dependent on the body's spontaneous activity, its material messiness, and indeterminacy. Hence, from the scientific perspective of causality, our freedom ultimately appears as groundless, as nothing but noise. Conversely, from a subjective perspective, our freedom can sometimes be felt as alienating, such during existential self-reflection or pathologically in schizophrenia. When we speak or think, the appropriate mental contents normally arise spontaneously into our awareness and organize themselves around our intentions in a way that is outside of our direct control. And even if we try to reflectively grasp at their origins, we cannot know exactly how or from where these mental contents bubble up into our stream of consciousness. This existential predicament is nicely captured by the robots' reports of their dawning sense of self-consciousness:

2. Robot: We have obtained a soul.

4. Robot: There is something in struggle with us. There are moments when something enters into us. We receive thoughts which are not our own.

There is an additional essential consideration about this existential mind-body gap that brings in a social dimension. The fact that there is an irreducible element of alterity in self-consciousness, that our embodiment prevents our consciousness from being completely transparent and closed within itself, is conversely what opens us up to the possibility of encountering the subjective perspective of others. On this view, our incapacity for complete self-transparence is not an anomaly of consciousness, but rather the flipside of our capacity to participate in each other's unfolding experiences, of our potential for genuine intersubjectivity.

We can find a hint of this philosophical insight in Busman's diagnosis that the robot's capacity to develop conscious awareness would not have been a problem if it were not for the fact that they interacted with each other in increasing numbers. There is a sense in the play that the dynamics of this social dimension have a life of their own, going so far as to prefigure the famous sandpile model of self-organized criticality according to which avalanche events follow a scale-free distribution: in such a system configuration, even a microscopic event can have macroscopic consequences at the collective level.

Dr. Gall: I only performed a number of experiments, no more than a few hundred.

Busman: […] This means that out of a million old, properly functioning robots just one will have been one of Gall's reformed models. Do you see what I mean?

Domin: So that means …

Busman: … that is has practically no significance at all.

Fabry: Busman is right.

Busman: I think I am. And now, lads, do you know what really caused all this to happen?

Fabry: What?

Busman: The number of them. We made too many robots.

[…]

Busman: Meanwhile, things gathered their own momentum, getting faster and faster and faster. Every miserable, greedy, dirty new order added its own pebble to the avalanche.

In summary, it seems reasonable that the subjective dimension of consciousness cannot be reduced to the objective body, but that it nevertheless expresses itself, indirectly and unexpectedly, in the gaps that animate the living body. Yet this concession is just the starting point: over time a human body only becomes a person by developing ways to channel this unruliness into the body's latent capacities, in particular by shaping and scaffolding them in social interaction. On the one hand, this irreducible interdependence between our subjective and objective embodiment implies that the reproduction of consciousness falls into the domain of biology, as Alquist realizes when he exclaims: *"If you want to live you'll have to breed, like animals!*". The key philosophical insight here is that in practice life can only come from life.

On the other hand, there is the important qualification that for this animal life to develop its full potential of consciousness, it must be situated in an appropriate social setting. It is this profound vision of consciousness as an organic phenomenon that only unfolds with the caring intertwining of self and others that is expressed in the play's poetic conclusion:

Alquist: […] life will not perish! Life begins anew, it begins naked and small and comes from love […] love, you flourish in the ruins, sow the seeds of life in the wind.