IRIE
International Review of Information Ethics Vol. 34 (10/2024)

Author(s): Walter Barta

The Dream of the Universal Constructor

Abstract:

Imagine inventing a "Universal Constructor", a machine that can molecularly assemble anything you want it to (within physical possibility). Should you construct such a machine? And what should you construct with it?

Agenda:

Author(s):

Walter Barta:

- Digital Research Commons, M.D. Anderson Library, University of Houston, 4333 University Drive, Houston, TX 77204-2000
- ✉ wjbarta2@uh.edu, 💻 https://libraries.uh.edu/drc/

ISSN 1614-1687

Walter Barta is a principal investigator at the Digital Research Commons at the M.D. Anderson Library at University of Houston. He is also a recurring contributor to the Popular Culture & Philosophy book series with a special interest in science fiction and philosophy.

Case Description

Imagine that by 20XX technological capabilities have become so advanced that machines can do nearly anything, and a team of scientists is on the cusp of inventing a "universal constructor", an object that, when instructed, can construct any possible object, including itself (von Neumann, 1966). The device works just like the microwave-like "Replicator" that the crew of the Enterprise in Star Trek use to prepare (or molecularly assemble) food (Saadia, 2016). But far from being limited to food preparation alone, running a series of tests, the scientists determine that the device will truly be universal in that it will be able to construct anything, including but not limited to tea and crumpets, top hats, tiddlywinks, spaceships, unicorns, elixirs of life, human brains, and neutron bombs. Of course, the device is limited by physical possibilities; it cannot produce anything nonphysical. Furthermore, the device is limited by conceptual possibilities; it cannot produce anything except what it has been instructed to. But even within these parameters is a design space filled with enumerable kinds of constructions, both terrific and terrifying. So, knowing that the world is full of potential beneficiaries who would enjoy the device and full of potential malefactors who would misuse the device, the scientists sit down for a serious meeting and discuss: once they can construct anything, what do they construct?

Questions

1. Is a Universal Constructor even possible? What further physical constraints might there be on a Universal Constructor? What physical possibilities would there be? What conceptual possibilities would there be?

2. Should we ever create a Universal Constructor? If so, how could we control its operation? If not, how could we prevent its creation?

3. Are technological affordances intrinsically good? Or are some affordances intrinsically bad?

4. Who should be allowed to operate a Universal Constructor? Why should they be given the privilege? Why should others be denied the privilege?

5. What should be allowed to be produced by a Universal Constructor? What should not be produced? Why should we create some things and not others?

6. Given the invention of a Universal Constructor, what do we produce first? What do we produce subsequently, and in what sequence? What do we produce last (if there is a last)?

7. What principles of use would be established for the Universal Constructor?

Exercises

1. Suppose that you could foresee the trajectory of technological progress all the way to its very end. What would that end probably be? What should that end ethically be? How would one nudge what would be towards what should be? Draw a trajectory of possible future technology to the end (or ends) you imagine. Include alternate paths for what could be or should be depending on how things might be nudged.

2. Suppose you had the idea for a Universal Constructor and knew it could be built within one year. As the inventor, what would you do to prepare? Draw up a to-do list of the things you would do to prepare. What would you add if the time span increased to ten years? One hundred years? Be prepared to justify them in discussion.

3. The physical constraints of the Universal Constructor could change the regulatory policies applied to its operating practices. For example, if the constructor could make uranium, then the regulatory policies around nuclear material would apply. List a few different possible physical constraints and the corresponding differences in regulatory policies. Of these constraints, choose at least one that you find particularly important and explain why.

Applying the Principles of AI Ethics

Using the chart provided, identify which principles of AI ethics are at issue in this case and, if principles conflict, which seems to be the weightiest and so the one that should override other principles.

Principle	Application (If Any)
Nonmaleficence	
Beneficence	
Respect for Autonomy	
Justice	
Explicability	
Accountability	

Normative Theories

Apply the different normative theories explained in the primer at the beginning of this volume to bring out issues in the case that might have been overlooked. How would a utilitarian approach this case? A deontologist? A virtue ethicist? Do any of these approaches accord with your own moral judgments or not? Do they arrive at the same verdict as the principles of AI ethics?

Expert Analysis (Read After Doing Your Own Analysis!)

In the 1940s, John von Neumann, the renowned game theorist and computer scientist, imagined the ultimate machine, the universal constructor, an object capable of constructing any including itself (von Neumann, 1966). Whereas von Neumann was interested in the logical, mathematical, and computational requirements of self-reproducing systems, other scientists have since taken up the concept and applied it to actual physical systems

in the real world. To this end, Eric Drexler imagined "molecular assemblers," nanotechnological devices capable of sorting and binding molecules, that humanity would eventually be able to use to produce almost anything under the sun (Drexler, 1986). Since then, various thinkers have dreamed about possible futures in which the operators of such devices have vast (nearly godlike) technological powers. As Nigel Calder (quoting Theodore Taylor) puts it: "Once such a machine exists it could gather sunlight and materials that it's sitting on and produce on call whatever product anybody wants to name" (Freitas, 2004).

In addition to the "Replicator" machine used in Star Trek, versions of the Universal Constructor have appeared in the worlds of numerous science fiction authors (Saadia, 2016). In Ursula Leguin's The Lathe of Heaven, a machine called the "Augmentor" is capable of turning dreams into reality: it can output any possible state of affairs from the input of a human thought. Similarly, in Michael Crichton's Sphere, such device is featured, the titular "sphere," which is speculated to be the technological core of a future, posthuman utopia.

Because dreaming of such a device can seem naively fantastical in its ambition, some writers, like Taylor, have derisively dubbed the Universal Constructor a "Santa Clause Machine" (Freitas, 2004). Others, like Manu Saadia, have more generously called the Universal Constructor a "metaphor for the distant endpoint of the Industrial Revolution" (Saadia, 2016). The device need not be just a metaphor though. In as much as the effect of technology is to afford its users new abilities, and technological invention ever increases such affordances, human progress is gradually moving slightly closer toward this dream (Norman, 2013). As technological affordances accrue and accumulate, the maximization of affordances in something approaching a Universal Constructor seems the natural, eventual endgame of history (Barrat, 2013; Bostrom, 2013). Indeed, ambitious engineers are already attempting prototype Universal Constructors. Just to name one example, the RepRap (replicating rapid prototype) Project is attempting even now to design a 3D printer that can print its own components (Jones et al., 2011).

Of course, a Universal Constructor will only produce what is possible. So, it is important to consider that the engineering of such a device will encounter physical constraints. Thermodynamics will not be violated, so no perpetual motion machines: neither can it 1) create energy from nothing, 2) nor perfectly convert heat to work. Relativity will not be violated either: neither can it 1) produce a superluminal (faster than light) spaceship, nor 2) a time machine to visit the past, nor 3) a big bang given finite input energies. It is also important to consider the logical constraints of the device's instructions. The instructions will have to preserve noncontradiction and there will have to be enough clarity to be interpreted correctly and uniquely. Furthermore, more specific constraints would surely exist as well, relating to technical factors in the device's design. However, even with constraints in mind, in the remaining design space of such a device there is still plentiful room for possible use-cases.

A merely quasi-Universal Constructor could pop out both possible utopias and possible dystopias willy-nilly. Thus, costs and benefits seem to be of primary ethical consideration when thinking about the designing such a device, which we might analyse using a roughly utilitarian analysis. To this end, the first important observation is that, because a Universal Constructor is (by definition) capable of maximal affordances, it is capable of both maximal beneficence and maximal maleficence (Norman, 2013).

On the one hand, by stipulation, such a device can afford any possible beneficence. It can alleviate every human necessity: all the food, all the water, all the shelter, all the medicine. Humanity need never need again. Furthermore, it can produce windfalls of wild potential, profuse and lavish luxuries, enabling the construction of an array of wondrous states of affairs, allowing users to fulfil every need or want with maximal efficiency. It could make spaceships a dime a dozen. It could make human lifespans as long as that of universe. It could make ambrosia-like delicacies never tasted before. It could make human beings as powerful and as happy as gods. Indeed, one might very well bring about what Arthur C. Clarke imagines to be a "utopia of infinite riches" (Clarke, 1977).

On the other hand, also by stipulation, if such a device constructs any states of affairs, it can also afford any possible maleficence. It could exacerbate every human misery: famine, pestilence, weaponry, and death. It could produce horrific pitfalls hitherto unrealized before except in philosophical thought experiment: brains in vats, evil demons, and more. Furthermore, any machine capable of near universal construction would also

(seemingly by necessity) be capable of near universal destruction. Such a doomsday device might irretrievably eliminate life on earth. Many have worried that a sufficiently sophisticated enough machine, if capable of replicating itself, might run amuck by converting the matter of the earth into copies of itself—what has been dubbed the "gray goo" scenario (Drexler, 1986).

All this considered, the cost/benefit analysis is difficult because the windfalls and the pitfalls seem equally extreme, so where does the balance tip? Perhaps we should be more fearful than hopeful. Why? Whereas a Universal Constructor must be used conscientiously every time to maintain a utopia, it need be used poorly only once to plunge the world into dystopia and devastation, perhaps destroying the universal constructor itself and its users along with it. In other words, whereas benefit must be maintained, damage may be both easy and irreversible. In this sense, a Universal Constructors make the world maximally wonderful but also maximally vulnerable (Bostrom, 2019).

Because the possible outcomes differ so drastically, appropriately assigning the responsibility and privilege of operating the Universal Constructor is of crucial importance. Thus, the character of the operator seems to be the natural next ethical consideration when thinking about using such a device, a consideration we might approach from a virtue ethics framework. Because the Universal Constructor affords maximal capabilities, we would want to assure that its operators would be maximally virtuous (and minimally vicious) in using it. Specifically, the operators should exercise the virtues complementary to their role, generosity in bestowing windfalls but cautiousness in preventing dangers.

With this in mind, one might be tempted to say that humanity should never produce a Universal Constructor, in spite of its potential benefits, because there is simply no human virtuous enough to be given the responsibility of operator—perhaps pessimistic, but fair enough. Realistically though, humanity may not have a choice but to create such a device, if the technological dilemma is that roughly that described by Francis Fukuyama: history's trajectory is directional because the acquisition of knowledge does not easily reverse, and knowledge once acquired confers decisive advantage upon those who acquire it (Fukuyama, 2012). So, once we discover how to make a Universal Constructor, we may not be able to undiscover that knowledge, and operators of the device may not be able to be stopped except by other operators.

Unfortunately, if true, this also means that the privilege to be the Universal Constructor's operator will perhaps be an arbitrary historical contingency, not an official role appointed to a maximally virtuous person. Whoever invents the device will be the first operator to wield it, and it is likely that this first operator will be able to disable other operators and defend themselves by using the device. Given this, the first operator may well be the only operator. This may be a lucky businessman, a disinterested scientist, or perhaps the corrupt president of some country or other. Whoever the operator is, it is likely they will not be maximally vicious, but it is also unlikely they will be maximally virtuous and may end up corrupted by their absolute power.

In order to avoid the tyranny of the inventor, we will be tempted to democratize the device by mandating that the first Universal Constructor be used to immediately construct a second (and third, and fourth, …), but universal democratization seems an unwise option as well, because it increases the probability of one of the devices falling into the hands of an unvirtuous agent, which seems like a worst-case scenario. Thus, it may be best to limit the number of constructors and operators to the smallest possible number that is still sufficient for beneficent outcomes.

One necessary precautionary measure would be for global human civilization—so that no affected parties are left out—long before the Universal Constructor's invention—before the privilege of usage is too late to revoke—to discuss and assign responsibility so that the eventual operators are indeed as virtuous as possible. This discussion would include the establishment of policies of development and usage and mechanisms of governance. This would include the regulation of activities of development and usage, oversight over the individuals and institutions developing the constructor, so as to know when intercession becomes necessary, and criminal prohibitions of risky activities, if necessary—not unlike current approaches to artificial intelligence policy (AIDA, 2022). How to moderate a negotiation with such high stakes amongst diverse parties of the world and with sufficient foresight would itself be a complicated issue.

But assuming the Universal Constructor will be invented, because of the further dangers of unvirtuous users, the first usage of the device would probably have to be the construction of an uncrackable security system around the device itself, to protect against potential misuse. Even so, the potential for unvirtuous (perhaps merely incompetent) operator error would be large, so the second usage of the device would probably have to be the construction of an unflappable quality control system to check and double check any instructions given to the device to prevent user error, a series of defensive systems for the physical protection of the constructor and its operator, and the full encryption of the design and function of the device to dispel attempts at hacking. Only then could one feel safe enough to construct Clarke's "utopia of infinite riches." Once finished with a stable and satisfactory world though, it might be safest for the final construction of the device to be a self-destruct system, disabling anyone from using it (or any other such devices) ever again, so as not to further enable unvirtuous misuse. However, since re-invention may always be physically possible, and perhaps politically uncontrollable (depending upon the powers afforded by the invention), it is not clear that such precautions would be possible.

In conclusion, although the Universal Constructor may itself never be constructed and may remain a dream, the possibility alone should be regarded with maximum hope and fear, for its potential beneficence and maleficence, and thus prepared for with the as much virtue as humanity can muster.

Reflection

Did you touch on everything this expert analysis identifies in your own analysis of the case? What did you miss?

Did you think of anything that could be added to the analysis or were there any points of disagreement? Be sure to justify your response with reasons.

References

Barrat, James (2013). Our Final Invention: Artificial Intelligence and the End of the Human Era. First edition. New York, Thomas Dunne Books.

Bostrom, Nick (2013). "The Future of Humanity," New Waves in Philosophy of Technology, eds. Evan Selinger & Soren Riis (Palgrave Macmillan, 2009): 186–215.

Bostrom, Nick (2019). "The Vulnerable World Hypothesis", Global Policy, Vol. 10, No. 3 (2019): pp. 445–476.

Calder, Nigel (1978). Spaceships of the Mind. New York, Viking Press.

Clark, Aurthur C. (1977). An Inquiry into the Limits of the Possible. New York, Harper & Row, Publishers, Inc.

Crichton, Michael (1987). Sphere. New York, Knopf.

Deutsch, David (2011). The Beginning of Infinity: Explanations That Transform the World. New York, Viking.

Drexler, Eric (1986). Engines of Creation. Washington, Fourth Estate.

Freitas, Robert A. Jr.; Merkle, Ralph C. (2004). "3.10: The Santa Claus Machine," from Kinematic Self-Replicating Machines, https://www.molecularassembler.com/KSRM/3.10.htm.

Fukuyama, Francis (2012). The End of History and the Last Man. New York, Penguin Books.

Jones, R.; Haufe, P.; Sells, E.; Iravani, P.; Olliver, V.; Palmer, C.; Bowyer, A. (2011). "Reprap-- the replicating rapid prototyper". Robotica. 29 (1): 177–191.

Le Guin, Ursula (1971). The Lathe of Heaven. US, Charles Scribner's Sons.

Norman, D. A. (2013). The Design of Everyday Things. Cambridge, MIT Press.

Saadia, Manu (2016-09-08). "The Enduring Lessons of "Star Trek"". The New Yorker. Retrieved 2019-07-24 from https://www.newyorker.com/tech/annals-of-technology/the-enduring-lessons-of-star-trek.

"The Artificial Intelligence and Data Act (AIDA) – Companion document." Canada.ca. Retrieved 2023-03-13 from https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document#s6.

von Neumann, John; Burks, Arthur W. (1966). Theory of Self-Reproducing Automata. University of Illinois Press. Retrieved 2017-02-28 from https://cba.mit.edu/events/03.11.ASE/docs/VonNeumann.pdf.