Review of: "More Human Than All Too Human: Challenges in Machine Ethics for Humanity Becoming a Spacefaring Civilization"

Neil Rowe[1]

1 Naval Postgraduate School

Potential competing interests: No potential competing interests to declare.

The paper seems to cover many well-understood issues, and does not seem to have much new to say since there are too many quotations of well-known positions by philosophers. Also, the paper focuses too much on "hard AI" and assumes that there is important work ongoing on hard AI, which is not true; nearly all work on AI is weak AI because hard AI has little value. There is not a single citation in the paper of work on "hard AI", which should have clued the authors that hard AI does not exist.

Also, the reference to space travel in the title is not supported by hardly any mention in the paper. Rather than space travel, it would make sense to focus the paper on just the limits of AI since that is much more what the paper discusses. It is unclear that AI is needed for space travel, since most of the technology for space is standard engineering. AI is only clearly needed in autonomous exploration of space.

Introduction in general: This fails to distinguish AI from other kinds of software. Most everything you say in this section applies to other kinds of software, so why talk about AI here? It is unclear that AI is needed for space travel. Space travel primarily needs to provide air, food and water, and energy, and those things do not require AI to obtain.

Introduction, second paragraph, first sentence: This is controversial and you need to give good arguments for it. Other planets in our solar system are not hospitable to human life and will require extensive infrastructure to use. Other star systems require hundreds of years

to reach by space travel. It would make more sense for humans to stay on their home planet.

Introduction, second paragraph: Define "normative" since this has many possible meanings.

Third paragraph: This issue does not make sense because people create AI and ethics applies to people and their artifacts.

"Defining artificial intelligence", first paragraph: The distinction between strong and weak AI is not useful in this paper because space travel is an engineering problem and can only use weak AI. Strong AI is purely speculative today and has not produced anything useful. The paper should be refocused on weak AI.

Third paragraph: Avoid pompous language like "inform the trajectory" instead of "affect". Pompous language hurts your persuasive ability.

Second and third sentences: These seem to contradict one another since they give very different definitions of SCOT. The first is commonly accepted, but the second is the opposite of the usual definition of science. You need to give more detailed arguments if you want anyone to believe the latter. The quote from Maze uses too much jargon to be easily understood by most readers and needs further explanation.

Fourth paragraph: You never define "technological determinism" and need to here. The "premise" you state is unlikely to be believed by more than a small minority because technology provides tools, and people design tools.

Fifth paragraph: Definition of what? You never say. Haugeland's quote expresses the opinion of only a small minority since machines with minds are not very useful for anything when we already have plenty of minds on our planet. You should acknowledge that this is a minority opinion.

"Ethical-impact agents", last paragraph: This needs to be expanded to be more precise to distinguish which AI methods are consistent with being programmed and which involve following an ethical principle. Neural networks would seem to the first, and rule-based systems would seem to be the second. Decision trees would be somewhere in between.

Rawlsian ethics would seem to be feasible to put into machines as machines can simulate possible worlds, and this could be considered rule-based.

"Can machines ever be full ethical agents?", second paragraph: It makes no sense to consider whether machines could ever be fully ethical agents because no part of society wants that for the reasons given earlier in the paper. So this is attacking a straw man. We always want humans in the loop for controlling agents, or else there wouldn't be much assurance they would serve our needs, and hence no reason to create them.

Third paragraph: The question is not difficult to answer because humans are machines. Their neural circuitry is a form of machine that operates within a finite set of resources. There's no biological evidence for a soul.

Tenth paragraph (counting the quotes): Here the paper detours into the dreary issue of consciousness, which has no relevance to the main topics of the paper. With weak AI, the consciousness of the creator is sufficient to provide the consciousness for an ethical theory embedded in the machine; the machine does not have consciousness, only needs to follow rules or criteria laid down by the creator, and these can be formulated to follow commonly accepted ethical theories. For instance, a mostly automated machine gun can be directed to fire only at people wearing military uniforms, and thus can follow the Geneva Conventions and the ethical theories behind them. In general, consciousness is just the higher-level process running brains, so there's no reason to impute mystical properties to it.

13th paragraph: Too many unnecessary vague terms are being introduced, like "mind" here. Why can't "consciousness" introduced previously suffice? There is no "mind-body problem" with weak AI because a machine is just running a program.

6th paragraph from the end: This is a pretty ridiculous-sounding argument since animals without brains are dead. The paper should stick to mainstream ideas if it is to make any useful progress.

"Should moral theory be extended to machines?", first paragraph: It

doesn't make sense to use the term "develop" for a machine that follows the principles of virtue ethics because machines can be programmed by a human programmer that understands virtue ethics. Same thing for utilitarian ethics.

Same place: The discussion of how moral theory should be extended to machines needs details. What sort of AI framework is envisioned? Something like a rule-based hypervisor, or will a neural network suffice? Can we assume explanation capabilities will be important to provide for ethical judgments? If so, what AI mechanisms are envisioned to accomplish them?

Sixth paragraph: This should mention the well-discussed issue of moral theory for weapons use in warfare. The weapons are not intelligent, but still many moral issues are addressed by militaries. Why would partly intelligent agents be any different?

"I, Robot", first sentence: This is highly arguable, but has no citation. Who believes such odd ideas? "Mental functioning and intelligence" occur in bacteria, but bacteria don't have minds. Then the second half of the sentence starts talking about human beings for no reason at all; animals can recognize mental states in other animals too. This sentence needs a rewrite.

Second sentence: No evidence for this applying to cognition has been presented in the paper, since many animals are able to think without communicating with other animals.

Third sentence: Which "subjective phenomena"? None are clearly identified. If you mean cognition and social interaction, that's what both psychology and AI have made major progress on in the last 100 years, and most of the principles of cognition and social interaction have been figured out now, so it's clear we understand the "physical mechanism" already.

Second through fourth paragraphs: It's unlikely that even strong AI will feel a need for self-preservation. Its "self" will be in the form of digital information which can be easily duplicated, so it can be brought back to existence repeatedly when its physical manifestation is destroyed. So strong AIs will almost always be of

less value than the lives of human beings, and should not need to be programmed with a high value on self-preservation.

Last paragraph: It's unreasonable to give even strong AIs the fundamental needs like that humans have (and certainly unreasonable for weak AIs). AIs are supposed to be servants and cannot do their jobs properly if they have many needs not of their masters.