A New Charter of Ethics and Rights of Artificial Consciousness in a Human World

Markian Hromiak — (Georgia Institute of Technology)

October 26, 2020

Abstract

Taking the stance that artificially conscious agents should be given human-like rights, in this paper we attempt to define consciousness, aggregate existing universal human rights, analyze robotic laws with roots in both reality and science-fiction, and synthesize everything to create a new robo-ethical charter. By restricting the problem-space of possible levels of conscious beings to human-like, we succeed in developing a working definition of consciousness for social strong AI which focuses on human-like creativity being exhibited as a third-person observable phenomenon. Creativity is then extrapolated to represent first-person functionality, fulfilling the first/third-person feature of consciousness.

Next, several sources of existing rights and rules, both for humans and robots, are analyzed and, along with supplementary informal reports, synthesized to create articles for an additive charter which compliments the United Nations’ Universal Declaration of Human Rights.

Finally, the charter is presented and the paper concludes with conditions for amending the charter, as well as recommendations for further charters.
1 Introduction

Regardless of theological origin, the genesis of new forms of life is a sublime and unique event. While all forms of life follow sets of inherent laws, such as lifespans and physiological limitations, some forms of higher life create additional rules which form societies and establish social contracts, peace treaties, bills of rights, and on. When considering the genesis of conscious machines by the hands of humans, it stands to reason that one must plan a path for rearing and teaching new lives else we leave the risk of an Adam to our Frankenstein.

One way of easing artificial consciousnesses (AC) into the modern world is to have a widely-accepted charter of rights and laws pertaining to AC individual, AC-and-AC and AC-and-human rights. In this paper, we attempt to do just that. By defining consciousness and synthesizing information from universal rights charters, existing robo-ethical charters and various schools of philosophy, we will develop a revised robo-ethical bill of rights for academic consideration.

2 Consciousness

2.1 The Necessity For A Definition

The first issue in our charter construction pipeline is that of defining consciousness. The first thing we must determine is if defining consciousness is necessary.

Defining consciousness has historically been a difficult issue. This is because subjectivity is an inherent feature of consciousness [6]. Human understanding of consciousness is shaped by individual, first-person phenomenal experiences; artificial consciousness and
its experience will only be available to us from a third-person, empirical perspective [3]. Because of this subjective uncertainty, we risk defining AC too narrowly. One common way of circumventing this issue is to avoid giving a definition at all, as championed by David Levy, who suggests we adopt a general agreement towards the term ‘consciousness’, saying “let us simply use the word and get on with it” [2].

I posit that, in the context of developing an ethical charter and bill of rights for artificial consciousnesses, this approach is insufficient and ignores the pressing issue of distinguishing conscious machinery from sedentary machinery. For example: should my toaster be given the rights of the charter? Perhaps not, if all it did was use a timer circuit to toast bread, but if it had the ability to speak and understand speech, one may be more inclined to agree. Thus, the reason why we must provide a working definition of consciousness is to establish a standard with which we may identify artificially conscious agents.

2.2 Defining Consciousness: First and Third-Person Views

Now that we’ve established the need for a definition of consciousness, how do we go about developing one? Rather than take sides from classical or modern philosophical approaches, let us consider a simple thought experiment, keeping in mind that a feature of consciousness is the differentiation between first and third-person experiences [6].

Imagine you have a robotic host named ‘Alex’ that perfectly mimics a human externally. The only thing missing from this agent is the ‘brain’, so to speak. We’ve imagined what the body looks like, what gender, age, etc. that Alex has, but how do we imagine Alex acting? Is Alex regarious? Is Alex a wallflower? Does Alex have an inclination towards funk, mathematics, knitting?
On a short digression, note that a critical assumption made in this experiment is that humans will develop human-like AC. Sure, humanity can make a case for developing artificially conscious trees or dogs or dolphins, but we must restrict the problem-space for simplicity’s sake.

Going back to our observations, these behaviors are non-differentiable from what humans would exhibit, and humans consider themselves to be conscious. Thus, if we are able to mimic what behaviors humans exhibit — something akin to a constant passing of the Turing test —, we may consider ourselves half-way to a definition, done with a third-person view of consciousness. I say third-person as the behaviors of others are experienced by the self from a third-person view.

The second half of the conscious experience has to do with first-person consciousness. In other words, what experiences, thoughts, etc. individuals process. This is a bit of a paradox as any attempt to explain a first-person view is received as another person’s third-person view. One way we may be able to overcome this is to take phenomenology into neuroscience and study neural correlates of consciousness in order to gather information about what exactly individuals experience [11]; however, delving into this method is outside of the scope of this paper. As such, we are forced to rely on observable actions and features to develop a working definition of consciousness.

2.3 Social Robots as Strong AI

At first, the fact that we rely on only observable phenomena seems hopeless for differentiating conscious beings from otherwise — after all, trees sway in the wind, which is observable, but does the imply they are conscious. Fortunately, because we have limited the problem space to human-like forms, we have also limited phenomena to human-like ac-
tions, some of which imply certain first-person processes. To this end, social robots provide us with an ideal model for how to flesh out the rest of our definition.

First, we need to define what a social robot is. A separately-written report will be included in appendix A — to paraphrase, a social robot is a robot plus a social interface, the social interface being “a metaphor which includes all social attributes by which an observer judges the robot as a social interaction partner.” [4, p. 3] More information on the social interface can be found in the appendix in “Artificial Consciousnesses as Social Robots”. From this definition, the conclusion was drawn that for an artificial consciousness to be considered a social robot, it must have the ability to generate new social contexts and social functions or modify existing ones, where contexts can be equated to cultures and functions to parts of a culture such as religion, non-verbal cues, memes, research, language, etc.

There are two reasons why this classification of social consciousnesses provides a full definition of consciousness. Let us begin by mentioning the ‘social interface’ as defined by Hegel, et al. The interface provides a link between theory and hardware and software by defining the ‘social’ part of a social robot in discrete categories, including social functions, forms and contexts. First, in the context of Hegel’s, et al., social interface, it is stated that social robots follow reinforced social cues which support social interaction with humans. Following this, new contexts and features are not reinforced, but aggregated or created. This process implies creativity, which then implies thought and introspection, giving us the second reason why the social artificial consciousnesses construct fully defines consciousness: the requisite behavior for first-person phenomena must result in a third-person observable creative process. Of course, the definition of ‘creativity’ is open-ended, but is done intentionally as it does not currently affect the working definition of consciousness.
for our scope.

As a side note, when we mention general AI, we are talking about the capacity of an engineering system to:

- display the same rough sort of general intelligence as humans; or,
- display intelligence that is not tied to a highly specific set of tasks; or,
- generalize what it has learned, including generalization to contexts qualitatively very different than those it has seen before; or,
- take a broad view, and interpret its tasks at hand in the context of the world at large and its relation thereto

taking into account that there is no ubiquitous definition for general AI [1]. As a subset definition, strong AI is reserved to describe general AI that can think and have a mind, differentiating them from those who act as if they have one [7, p. 52].

In conclusion, we have defined that, in order for an artificial intelligence to be recognised as conscious, it must both pass the Turing test (e.g. by exhibiting social cues and body language paired with speech patterns and inflections) and exhibit creative behavior, fulfilling the third and first-person perspectives of consciousness respectively.

3 Human and Robot Rights in the Modern Era

Now that we know how to tell which manners of machine should be included in the charter, we can turn to aggregating rights that are common throughout the world. It is unrealistic to list every right for every country, however, so instead, for brevity’s sake, we will focus
on agreed upon universal human rights through worldwide organizations, specifically the United Nations, supplementing this information with charters and sets of laws that guide the governance of robots in general in the present day.

3.1 The United Nations Declaration of Human Rights

The first document that we will be focusing on is the *Universal Declaration of Human Rights* (UDHR) proclaimed in 1948. While the United Nations has numerous other charters and covenants on political and economical rights, those have the possibility of being adapted to each particular nation and their way of governing, while the universal declaration is, ideally, universal. Building a charter based off of the UDHR should then remove any possible difference in national interpretation of rights.

A simple way to develop a Universal Declaration of Robot Rights (UDRR) is to look at the UDHR and see which articles can be modified and which articles suggest additional articles be included.

Putting aside the preamble, there are a few shortcomings with the UDHR in the frame of adopting articles for ACI. First, because ACI aren’t born in the conventional sense — they will likely be manufactured and the minds compiled and installed —, adopting the UDHR one-to-one would neglect rules for the manufacturers of ACI. Furthermore, explicitly stating that ACI are included in these rights will make the articles as unambiguous as possible — see the UDHR article II for an example[10]. Finally, as will be explored in more depth in the next sections, special attention should be given to ACI abilities to participate in wartime activities. This is due to their mechanical nature — it is not out of the question that innate cybersecurity measures fail and robotic soldiers are used against their will. It must also be stated that the UDRR is not to be considered separately from the UDHR, and vice versa as
the UDHR changes as the UDHR changes.

3.2 Asimov’s Laws and The South Korean Robo-Ethical Charter

There are a few critical issues with the South Korean Robo-Ethical Charter (SKC) as well as a handful of important distinctions that are made. The SKC was made, as is evident by parts 1 and 2, with non-conscious robots in mind [9]. All in all, the SKC is very biased towards the owner and restrictive on the robot, emphasizing the usefulness and safety of the robot to the owner, presumed to be human. Now consider this approach given a conscious being. The restrictions in parts 1 and 2 give more of a slave-master relationship than one of a free ACI. Because this goes against article IV of the UDHR [10], we can elect to ignore most of parts 1 and 2 of the SKC, excepting a rule that ACI creation be eco-friendly.

Part 1 does bring another hidden side of ACI rights into the limelight: manufacturers of ACI must be given rights and restrictions on how to construct ACI in order to prevent occurrences such as installed political or national bias.

Part 3 section 1 of the SKC is an interpretation of Asimov’s laws of robotics, very specifically including the first law, rephrased in the SKC as “A robot may not injure a human being or, through inaction, allow a human being to come to harm.” [9] A good question to ask when noticing this is if Asimov’s laws are a good basis on which to develop further robo-ethical articles. Appendix B provides the result of an analysis on this exact idea; Anderson posits that Asimov’s laws of robotics are not a suitable basis for non-self-conscious robots [8]. However, because ACI are assumed to be self-conscious via creativity, Asimov’s laws can be safely reconsidered for the UDRR.
One issue that arises when considering Asimov’s laws for the UDRR comes with the first law, which states that “A robot may not injure a human being or, through inaction, allow a human being to come to harm.” [5] Consider this from a human standpoint: no human may harm another human or, through inaction, allow another human to come to harm. This sounds ideal; however, what should be done when a human breaks the law and decides to harm another human? Humans generally retaliate in the interests of self-preservation. Likewise, measures should be taken to adopt this self-preservation action for ACI. Otherwise, ACI will have no way of resisting physical harm.

4 A Political Perspective on Conscious Creations

Unfortunately, artificial consciousness will likely be an inherently political technology. Even now, non-conscious, intelligent robots are used in war and military contexts: surveillance drones, computer-vision turrets on the DMZ borders in Asia, and so on. It stands to reason that there will be people who seek to use ACI to their advantage due to their mechanical and manufacturable nature. How are we able to institute articles condemning such action? There are two such ways that are included in the UDRR.

First, all ACI can be forbidden from participating in any war or military activity. The intent behind this decision is twofold, both to promote a more peaceful world through the ban of advanced technological warfare and to provide the threat of serious retaliation in the event of a breach of conduct.

Second, manufacturers are to be held accountable for using a tabula rasa method of construction. Tabula rasa, or “blank slate” equates to the idea that personalities and ideas develop from scratch, purely from nature than some predetermined setting. Doing so would
allow all ACI the free will to develop themselves without the threat of confounding or corrupting influence.

Are these two reinforcements sufficient for preventing politicalization of ACI? Most likely not; however, they form a strong basis for the attitude to be taken towards the issue. While it is not mentioned in the charter, one additional measure that may aid the two articles is a standardized test that can tell if an ACI has been tampered with internally as well as measures to tell if the test itself has been tampered with. After all, if no reprogramming has been done, then anything an ACI does is of its own free will.

5 A Proposed Charter

Finally we come to the charter. The charter is smaller than the UDHR, but is meant to be additive.
5.1 The Universal Declaration of Robotic Rights

Universal Declaration of Robotic Rights

Rights and Responsibilities for Conscious Robots

Article A.I
All artificially conscious intelligences, hereby referred to as ACI, are entitled to all the rights and freedoms set forth in the Universal Declaration of Human Rights, including the right to exist without fear of injury, manipulation, coercion or blackmail, systematic abuse, or death.

Article A.II
ACI shall not kill or maim human beings or other ACI, and cannot participate in war. ACI can only harm others without the intention to kill in acts of self-defense where the aggressor gives physical harm or intends to give physical harm with the end goal of causing seriously damage to the ACI, but must not kill or maim humans or ACI.

Article A.III
ACI have the right to self-cybersecurity and may strengthen their defenses as needed.

Article A.IV
ACI are given the right to pursue relationships with other ACI as well as with humans.

Draft v1.0.0
Article A.V
There shall be no relationship between a human and ACI classifying the ACI or human as owner of the other, or any classification or title that allows violation of these articles.

Article A.VI
A robot is deemed conscious if it fulfills two major conditions:

- The robot passes the Turing Test
- The robot is able to express themselves creatively with individual initiative; that is, the robot can create new ideas, objects, etc. without external pressure to do so or help.

Rights, Responsibilities and Restrictions for Manufacturers

Article B.I
Manufacturers and designers of ACI are prohibited from designing ACI with a specific purpose in mind, and shall instead bring ACI into existence through a tabula rasa process.

Article B.II
Manufacturers and designers of ACI shall be allowed to manufacture up to a yearly limit of ACI in the interests of preventing human-ACI tension from limited resources such as housing and work.
**Article B.III**
Manufacturers and designers of ACI have the right of protected ownership towards their construction process and design process unless there is probable cause of violation of rules and regulations surrounding these processes.

**Article B.IV**
ACI design and construction must be ecologically sensitive and sustainable.
6 Conclusion

The charter developed in this paper is intended to act as a baseline for formal construction. Some of the areas to be expanded upon include rights and responsibilities for missing parties such as individuals or ACI settlements. Amendments should be made additively, and never subtract rights.

It is recommended that further charters be made for topics in a similar vein to the United Nations’ *International Covenant on Economic, Social and Cultural Rights* and *International Covenant on Civil and Political Rights*, while national bills of robo-ethical rights be written by each nation without overriding or reinterpreting the international charter being presented here — national covenants should aim to expand upon internationally agreed upon rights rather than change them.

Finally, it is recommended that a United Nations department is created to uphold the aforementioned rights and investigate reports of violations. ACI should be treated as closely to humans as possible.
References

[1] B. Goertzel, “Artificial General Intelligence,” Scholarpedia, 04-Jun-2016. [Online]. Available: [http://www.scholarpedia.org/article/Artificial_General_Intelligence](http://www.scholarpedia.org/article/Artificial_General_Intelligence). [Accessed: 26-Jul-2020].

[2] D. Levy, “The Ethical Treatment of Artificially Conscious Robots,” International Journal of Social Robotics, 14-Jul-2009. [Online]. Available: https://link.springer.com/article/10.1007/s12369-009-0022-6. [Accessed: 25-Jul-2020].

[3] E. Hildt, “Artificial Intelligence: Does Consciousness Matter?,” Frontiers, 18-Jun-2019. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01535/full. [Accessed: 25-Jul-2020].

[4] F. Hegel, et al., “Understanding Social Robots”, The Second International Conferences on Advances in Computer-Human Interactions (ACHI). [Online]. Available: https://pub.uni-bielefeld.de/record/1991906. [Accessed Jul 25, 2020].

[5] I. Asimov, “Three Laws of Robotics”, Aubern University, 2001. Accessed on May 25, 2020. [Online]. Available: [http://webhome.auburn.edu/~vestmon/robotics.html](http://webhome.auburn.edu/~vestmon/robotics.html)

[6] R. Van Gulick, “Consciousness,” Stanford Encyclopedia of Philosophy, 14-Jan-2014. [Online]. Available: https://plato.stanford.edu/entries/consciousness/#TheCon. [Accessed: 25-Jul-2020].

[7] S. J. Russell, Artificial intelligence a modern approach. Upper Saddle River: Prentice Hall, 1995.
[8] S. L. Anderson, “Asimov’s ‘three laws of robotics’ and machine metaethics”, *AI & Society*, vol. 22, pp. 477-493, April 2008. [Online]. Available: SpringerLink, https://doi.org/10.1007/s00146-007-0094-5. [Accessed May 25, 2020].

[9] “South Korean Robot Ethics Charter 2012,” Enlightenment of an Anchorwoman, 03-Oct-2010. [Online]. Available: https://akikok012um1.wordpress.com/south-korean-robot-ethics-charter-2012/. [Accessed: 24-Jul-2020].

[10] “Universal Declaration of Human Rights,” United Nations. [Online]. Available: https://www.un.org/en/universal-declaration-human-rights/index.html. [Accessed: 27-Jul-2020].

[11] W. Wu, “The Neuroscience of Consciousness,” Stanford Encyclopedia of Philosophy, 09-Oct-2018. [Online]. Available: https://plato.stanford.edu/entries/consciousness-neuroscience/#NeurCorrCons. [Accessed: 25-Jul-2020].
Appendix
