Ethel the Ethical Robot
and Illustrating Dialogical Approaches to AI Programming
Gray Cox May 14, 2022

How might we begin to illustrate some underlying general principles that can help guide a human ecological path toward dialogical forms of AI technology? Since 2015 there has been rapidly increasing work on ethics in AI that has taken many different forms which are relevant to this. Two core principles that are emerging concern the goals of the processes involved and the the methods:

Strategy #1: Our guiding goal should be genuine, voluntary agreements.

&

Strategy #2: Our methods should be those of collaborative dialogue.

These two strategies are integrally connected. Each helps define and explain the other in ways that set them apart sharply from the Smarter Planet goals and methods. The monological forms of "rocket science" reasoning that drive the Smarter Planet vision of technology assume that the goal of reasoning is to arrive at conclusions using methods of inference that rely on algorithmic rules of formal logic. In contrast, technology for a Wiser Earth will be guided by the aim of generating knowledge, making decisions, and managing our world in accordance with agreements amongst all the participants who represent diverse points of view in the dialogue. And for those agreements to be genuine and voluntary, the participants need to negotiate them in the kinds of dialogical ways we have discussed earlier. They need to negotiate agreements about the meanings of their terms and background assumptions. They need to problem-solve for ways to escape dilemmas and create new, viable options that do a reasonable job of dealing with everyone's' interests and concerns. They need to resolve or transform conflicts in just, sustainable, resilient ways. Further, if the agreements are genuine and voluntary, they need to be uncoerced. Methods of nonviolence need to be used to discern, demonstrate, and defend claims for emergent, objective moral truths.

The relationships between the goals and the methods here are interconnected in a way Gandhi captured with a metaphor: "The means may be likened to a seed, the end to a tree; and there is just the same inviolable connection between the means and the end as there is between the seed and the tree." You cannot get a maple tree by planting an oak acorn. Likewise, you cannot get genuine, voluntary agreements by reasoning in monological ways. Such agreements can only be arrived at through dialogue that is grounded in nonviolence and framed by respect for each other and the larger environment that is rooted in emergent objective moral truths.

A corollary of this is that the computer systems involved need software that is treated as an agent rather than a tool and which is developed in ways that let it offer proposals for dialogical agreement rather than simply and solely generating conclusions from monological inferences. To do this, such agents need to be able to shift back and forth between running code and calling it in to question. They need to be able to engage with their own structure at meta-levels that permit them to reconfigure all of their elements in the process of renegotiating the assumptions, data,

values and other terms at play in the dialogue in which they are involved. As stressed before, this need not mean that the machines need some new super-duper futuristic form of hardware. Instead, it means that the structures of reasoning which they incorporate need to be dialogical rather than merely monological.

A rudimentary version of this can actually be introduced into very simple systems for programming like the SCRATCH language developed at MIT for teaching block coding to children. (The program is available here: https://scratch.mit.edu/projects/428374274.)

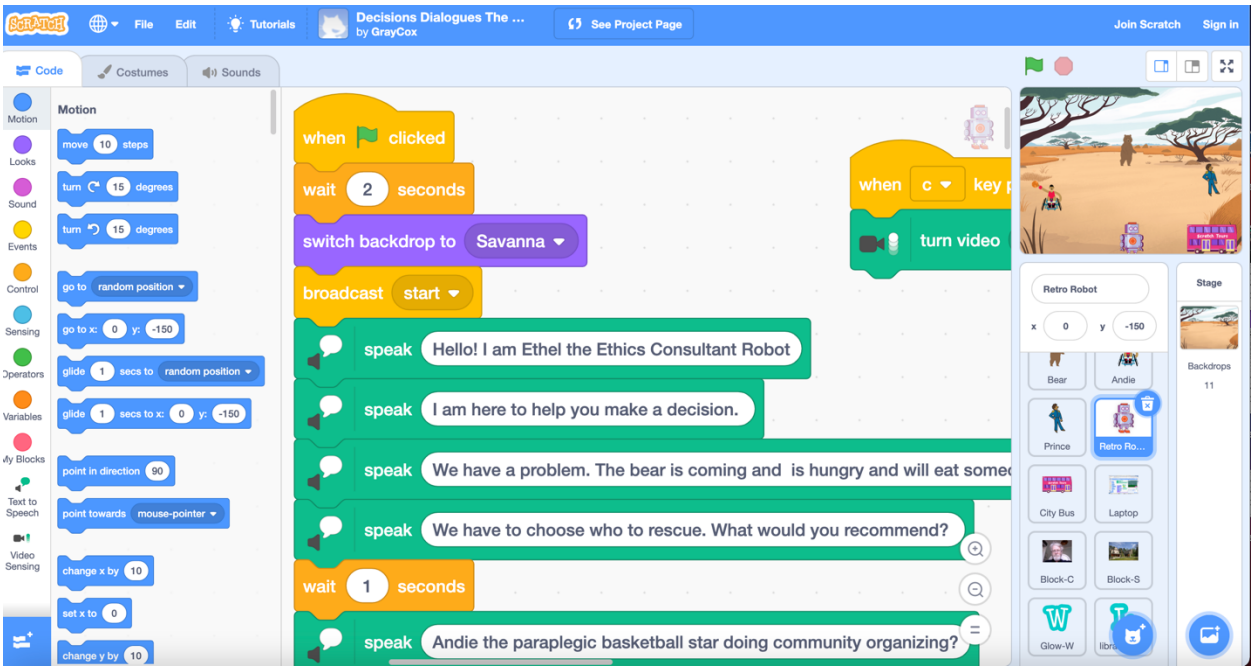
In it, the process of reprogramming is actually carried out by kids and others using the program, but the program is designed to prompt questions and dialogue that lead to helpful forms of reprogramming using principles of collaborative negotiation and shared problem solving.[i]

In one example of this, the character of Ethel the Ethical Consultant Robot offers to help kids wrestle with a decision which is initially presented in the form of a classic Trolley Car dilemma. They are in a wild open place with a tour bus and see two people out observing different parts of the savannah. One is a young black man in a wheelchair who is characterized as a paraplegic community organizer, the other is an older person characterized as a prince who does ecological restoration. A wild bear is approaching and threatens to eat them. Ethel indicates that there is only time to rescue one of them with the bus and goes on to walk the game player through some considerations comparing the two potential victims' importance to society, potential for future life, and their previous suffering in order to arrive at an ethical decision about which one to sacrifice. Thus far, the game makes use of classic styles of unilateral thinking

exemplified by Enlightenment ethics and demonstrates the kind of AI programming associated with the algorithmic reasoning associated with the standard model of Turing Machines. But then something a bit different starts to happen.

Ethel notes that it may have been hard to make the choice and invites the player to reflect on how and why – and how a better decision process might go in the future. She asks if there might be other interests or concerns that would be worth considering in making the decision. She asks if there might be some creative third options the player might suggest as alternatives to simply using the bus to rescue one person or the other. She keeps track of the player's answers and then feeds them back in a summary along with a suggestion. The suggestion is that the player go in to her code and revise it to include these new ideas. She also notes that she has been programmed to weight the relevance and importance of the various interests equally but perhaps the player would think this should be changed. Also, the player might think there are other features of her code that should be changed, and they are invited try this and talk with their class mates about it. Ethel closes with a general invitation to the players to adopt an open minded and collaborative approach to ethical thinking in which they may continue a process of reprogramming her and working with her in successive ways so that, through this kind of dialogue, they can all make some progress in becoming wiser.

The programming process for this uses a kind of block coding that is visually based. You just drag and click colored chunks of code on top of each other in more or less the way you might



build something with Legos. In the process, a player learns about the power of basic operators like “if then” and learns to manipulate the visual images and sounds that create the video imagery of the game. Still, in a very primitive but useful way, this program illustrates some of the key principles of collaborative reasoning in ethics and dialogical programming for AI. It does not, of course, by any stretch of the imagination count as real dialogue. All of the wight of the creative and critical problem solving is done by the player. The machine is just flagging some issues and providing some prompts for the player to reflect on and discuss with classmates – and use to reprogram “Ethel”. The kind of “meta-cognitive” and dialogical activity that this reprogramming provides a representation of is really quite limited. It relies only on symbolic logic functions of the kind found in Good Old-Fashioned AI and does not include any of the capacities of evolutionary, connectionist kinds of activity that have come to provide powerful pattern recognition and generative systems since 2012.

Much more sophisticated and systematic approaches to dialogical practices of programming are possible in languages like Python in which the meta level activities of dialogue can be structured and facilitated by websites like Github. These allow programmers to collaborate in writing code in ways that facilitate the processes of developing proposals, exploring their flaws and merits, and finding group consensus to adopt or commit to some branch of the possible revisions of the program. Many of the key challenges for advancing dialogue based, human ecological programming lie in finding ways to adapt and expand the ways in which: 1.) these forms of collaboration can be elaborated and extended to all the levels and aspects of the programs and machines involved in a system that is managing a school or farm, and 2.) these forms of collaboration can made transparent and accessible so that all the points of view, sources of wisdom and varieties of intelligence in those systems can take part in the process of assessing and revising the code. There have to be ways in which teachers and farmers who do not write in Python can still engage with the Github site that can benefit from their knowledge about how and when machines can make useful judgment calls and how and when skilled individuals or concerned groups need to play key roles in making decisions.