Morality in the Machine Age:
Keeping Humans
in the Loop

A Templeton World Charity Foundation Challenge

June 2017

Goal:

Human moral capacities need not be diminished by new technologies and the 'machine age' need not be a period of sinister, controlling algorithms. Our goal is to support new systems for human-machine interactions that will provide tools to empower human moral intelligence, enhance our ethical capacities, and propel human spiritual betterment.

Opportunity:

Humans and our technology exist in a complex feedback loop. As we create new devices, they, in turn, influence us, both enhancing and circumscribing our life experiences. This loop is particularly powerful in the case of artificial intelligence which amplifies our capacities but also offers the chance to abdicate some of our responsibilities, delegating choices to clever machines. TWCF will support research into this loop and the development of theories, models, and technologies that support human moral strengthening and advancement. We seek to support approaches that combat the possibility of humans becoming simply passive consumers of circumscribed choices generated by our own machine aids.

Research and development of artificial intelligence is richly funded by those seeking profit. Vast resources are currently deployed in search of 'monetizable' theories and technologies but there remains little direct funding focused on the how these innovations impact human agency, moral capacity and development. This presents an unparalleled opportunity to leverage our resources by partnering with laboratories, researchers, and companies that have other funding for their technical work but that seek additional co-funding to support specific research into and development of approaches to enhance human flourishing. We anticipate that the bulk of grants under this challenge, therefore, will be highly leveraged against funds designated for purely technical aims.

We already use AI in many contexts with real human impact, ranging from immigration control in airports to determining power distribution in towns and cities. These uses are increasingly joined by far more subtle applications, with machine logic sometimes substituting for judgments made by humans, decisions that can carry tremendous moral or ethical weight. Among the most celebrated of these are algorithms that help governments decide which criminals are the most likely to offend again, algorithms which the U.S. Attorney General recently worried might "inadvertently undermine our efforts to ensure individualized and equal justice". Ensuring just outcomes should be a prerequisite of all such applications but, because of the state of the art in AI, that can be challenging. With the economic efficiencies and predictive power that these algorithms provide, it will require diligence to ensure that moral human decision-making remains foregrounded – that it is fully embedded in and empowered by these new tools.

In some of the above cases, the machine age seems to allow us to 'outsource' our decision making. Of course, banal decisions such as what song will come up next on our playlist seem to lack moral or ethical content (but may, in fact have subtle and important impact). But in more momentous situations we certainly risk ceding part of our inner compass. How do we ensure that the instructions and values that we program into these machines are consistent with our own, or that the choices we ask them to make on our behalf will be consistent with a benevolent outlook on the future of human existence? And, in a world of many (sometimes conflicting) ethical systems, how can we best elucidate the structure of that decision making? These questions must be answered if we are to ensure that our innate capacities strengthen rather than atrophy.

Such questions have begun to be explored in the fields of computing, artificial intelligence, robotics and subfields of philosophy, theology and the social sciences. The nascent study of machine ethics has begun to ask the question of how to implement moral decision-making in computers and robots. Some researchers, for example, have already begun to use stories to "teach" human values to AI systems to help them perform morally-neutral actions. We will seek such points of leverage where novel approaches offer particular promise.

The most promising approaches intentionally embed an element of human reflection and interaction – they seek to prevent us from becoming morally flaccid by regularly querying us and seeking ongoing 'opt-in' or auditing. They try to combat the tendency of some of these systems to push human ethical decision making to the background of our social-technological interface. Deep human feedback is crucial if we are to remain fully "in the loop" -- exercising moral choice in the service of gaining ever greater moral and ethical capacity. We will seek especially to fund those approaches which focus particularly on this 'looping in' of human agency.

Such approaches highlight the very real possibility that we can learn to be better by creating AI tools that help us to more clearly see the moral and ethical quandaries we face and, perhaps, give voice to our own better angels. Because AI has the capacity to detect patterns and to perceive connections that often go unnoticed, such machines could, in principle, augment our own capacities and help us reflect, grow, and act better.

This sort of aid might be seen as an enhancement, a learning tool that extends rather than replaces our own perception, choice, and capacity. We challenge researchers to explore the possibility that, by dint of their capacity to digest much more information than humans can, for example, machine intelligences could expose conditions, situations, and opportunities that we might otherwise have missed, thereby enhancing our capacity for choice and moral, ethical action and attention to the most important aspects of life.

One of the grave challenges presented by all of these feedback loops is the opaque nature of some machine intelligences. Artificial neural networks, for example, function as closed 'black boxes'. They clearly have rules that associate inputs to outputs but current technologies do not offer human insight into what those rules actually are. The result is that we might build and rely upon machines to help us make moral decisions that only *seem* to be consonant with our expectations and ethical precepts. It would be valuable to gain deeper insight into the workings of such tools, to be able to *interpret* the rules that may be implicit in their inner working so that they become more transparent elements in the complex loops of human decision making and social change. Some research on this is currently being supported and TWCF will seek to co-fund projects in this area with particular foci on the moral and ethical contents of the inner programming of machine intelligences.

Terms like 'moral' and 'ethical' are laden with personal, cultural, philosophical, and theological content and work in this area should take a sophisticated approach to the subtle issues that arise. We seek to engage thinkers across geographic and disciplinary boundaries and to incorporate their insights into the technical work we support.

Roadblocks

Engineers are not ethicists, and ethicists (generally) do not know how to write computer code; we will need highly collaborative teams of experts from multiple disciplines to contribute to this challenge. A comprehensive and robust framework for human/machine morality is likely to require knowledge and expertise across typically-unconnected disciplines. A first step towards realizing the goal described in this challenge statement is to facilitate cross-disciplinary constructive dialogue.

There is also a need for disciplines involved in the conversation to re-think and push beyond current boundaries. For example, the study of morality and ethics has often attempted to boil highly complex situations down into simplified scenarios for the purpose of theoretical exploration. To consider how a human/machine feedback system should be involved in moral decisions in the real world, it will be necessary to have robust accounts of both the underlying principles that might guide actions (based on this sort of theoretical exploration) as well as how these play out in the real world and can be leveraged to develop future technologies and enhance the human-AI loop (a more practical ethics).

In addition, there is the roadblock of moral plurality. “Morality” or “ethics” are not singular monolithic entities agreed upon by all ethicists or moral teachers. Morals, values, and ethics exist across a wide spectrum of views, influenced by religion, cultural traditions, political trends, and circumstantial accidents. Even to draw out one principle as a “common thread”, such as the so-called Golden Rule, is not much help since the practical application of “as you would be done by” is depends to a significant extent on one’s worldview. Depending on the school of ethics one subscribes to (e.g., consequentialist, ontological, or otherwise), opposing actions could end up being considered the “most ethical”. Therefore, identifying and clarifying which ethical or moral framework to use - a philosophical rather than a scientific question – will be a key challenge.

From a technical point of view, there remain deep inconsistencies between the perceived world of humans and machine intelligences. On the one hand, machines have the capacity to ingest vast quantities of data in forms simply unavailable to human actors. And they can successfully build problem-solving strategies in very circumscribed situations far better than humans. On the other hand, humans naturally pre-process the information the world provides, characterizing facts and ascribing motives, for example, in such a way as to lay the groundwork for using those inputs to make decisions. Our general capacities, the ability to intuitively connect similar but different kinds of situation, and our flexibility of mind are far superior to any existing machine intelligences. Learning to fruitfully interleave these two very different sorts of ‘mind’ (and developing appropriate hardware and software) presents a major roadblock to developing morally enriching AI tools and techniques.

Challenge Statement

Develop theories, models and technologies that make it possible for the human use of AI tools to enhance moral capacities, enrich ethical action, and propel human spiritual betterment.

1. Craft a systematic overview of the ways in which moral capacities are reflected in, magnified or diminished by, and given new contexts for by the human-AI loop.
2. Build bridges between faith traditions and advancing AI research communities.
3. Enhance mature, maturing, or proof-of-concept stage AI research and development with distinctively moral/ethical foci.
    a. Attempt to allocate many of the grants under this challenge through a co-funding model.
4. Develop tools, techniques and processes to ensure that human agency remains foregrounded in actions of moral consequence involving artificial intelligence.
    a. Identify and investigate kinds and situations of ‘outsourcing’ of human moral capacity to machine intelligences.
    b. Explore methods by which machine intelligences can be made to act in consonance and/or partnership with human moral decision making.
    c. Identify and develop approaches to strengthening rather than atrophying human moral capacities in the human-AI loop.

    d. Engage new subfields and interdisciplinary groups with particular foci on machine ethics and novel methods to imbue machines with general capacities spanning complex situations in which moral and ethical issues can be located.
    e. Develop strong feedback elements to ensure active participation, reflection, auditing, opt-in, etc. by humans to fully engage them in the loop on morally significant action.
5. Use AI's unique capacities to broaden the horizons for human moral action and to provide new forms of information on which to base human decision making.
    a. Develop tools, aids, prostheses, etc. that strengthen human capacities especially in the context of well-established and ongoing moral commitments.
6. Investigate making artificial intelligence's internal workings interpretable so that *seemingly* moral decisions made by machines can be fully understood by humans.
7. Develop approaches to ensure clarity and openness in explicating the varieties of moral/ethical/philosophical frameworks upon which human-AI loops should be based. The Foundation does not expect its funded research to embrace any particular religious or spiritual framework.

Areas we do not fund

1. Research or development of fully autonomous machine intelligences.
2. Research or development of technologies that seek to replace human moral decision making with that of a machine.
3. Explorations of the moral/ethical status of machines themselves except as part of a broader exploration of agency and capacity that spans the wide varieties of intelligences.
4. Generic machine supervisors or AI morality monitors that do not operate within a specific context of ongoing human choice or a specific set of commitments.