On the Promotion of Safe and Socially Beneficial Artificial Intelligence

Seth D. Baum

Forthcoming, *AI & Society.*
This version dated 1 October 2016.
Preprint at: http://sethbaum.com/ac/fc_AI-Promotion.html

Background: Improving the Societal Impacts of Artificial Intelligence

As AI becomes more and more capable, its impacts on society are also getting larger. AI is now being used in medicine, transportation (self-driving cars), the military (drones), and many other sectors. The impacts of AI on society depend a lot on how the AI is designed. To improve AI design, two challenges must be met. There is the technical challenge of developing safe and beneficial technology designs, and there is the social challenge of ensuring that such designs are used. *This paper studies the social challenge of how to promote safe and socially beneficial AI designs (or simply "beneficial AI" for short).* The paper looks at both near-term and long-term AI, including the possibility of future AI being superintelligent: significantly more intelligent and more capable than humans. There has been little prior research on promoting beneficial AI, so the paper uses examples from other contexts that have been studied more, such as environmental protection.

Extrinsic Measures

Most efforts to promote beneficial AI focus on extrinsic measures, which are imposed on AI communities from the outside. Government regulations and institutional policies are some common forms of extrinsic measures. Extrinsic measures can be in the form of constraints, in which certain AI designs are required or forbidden, or in the form of incentives, in which certain AI designs are encouraged or discouraged. Extrinsic measures are generally built on the premise that AI designers do not want to choose beneficial designs—otherwise, the measures wouldn't be needed. Therefore, the measures generally require mechanisms for ensuring compliance and punishing noncompliance.

Intrinsic Aspects of Extrinsic Measures

Extrinsic measures cause reactions within AI communities. These reactions are the intrinsic aspects of extrinsic measures. The reactions can determine whether the extrinsic measure is successful at making AI more beneficial. Some extrinsic measures cause positive reactions, making them succeed with minimal enforcement. For example, laws requiring dog owners to cleanup after their dogs are generally successful even though they are rarely enforced. Other extrinsic measures cause negative reactions, making them less successful. Some extrinsic measures can even backfire, causing worse outcomes than if there was no measure. For example, it is sometimes believed that a constitutional amendment to ban flag burning would result in more people burning flags because it would turn flag burning into a protest against government repression of free speech. *In order for extrinsic measures for AI to succeed at benefiting society, it is important that they be designed in consideration of intrinsic aspects*.

Dedicated Intrinsic Measures

Beneficial AI can also be promoted via dedicated intrinsic measures, which motivate AI developers to want to pursue beneficial designs. There are several types of dedicated intrinsic measures. For example, social norms can be cultivated within AI communities such that they think they should pursue beneficial AI. Respected messengers can be used to deliver messages so that AI developers are more likely to listen. Messages can be framed in ways that make beneficial AI more appealing. Stigmas can be applied to AI that is not beneficial. These and other dedicated intrinsic measures can be a powerful way to promote beneficial AI.