Markkula Center
for Applied Ethics
at Santa Clara University

A Multilevel Framework for the AI Alignment Problem

Betty Hou ’22
2021-22 Hackworth Fellow at the Markkula Center for Applied Ethics

Brian Patrick Green
Director, Technology Ethics, Markkula Center for Applied Ethics

A Multilevel Framework for the AI Alignment Problem

Betty Hou, '22 *is a Santa Clara University Graduate in Computer Science and Engineering and was a 2021-2022 Hackworth Fellow at the Markkula Center for Applied Ethics at Santa Clara University.*
Brian Patrick Green *is the director of the technology ethics program area at the Markkula Center for Applied Ethics at Santa Clara University. Views are their own.*

Introduction: AI Ethics

"'You were going to kill that guy!' 'Of course. I'm a Terminator.'" Lines like this from the 1991 James Cameron film *Terminator 2: Judgment Day* presented a dark warning for powerful, malicious artificial intelligence (AI) [1]. While a cyborg assassin traveling back in time has not yet become a major concern for us, what has become apparent is the multitude of ways in which AI is used on a global scale, and with it, the risk of both direct and indirect negative effects on our political, economic, and social structures. From social media algorithms, to smart home devices, to semi-autonomous vehicles, AI has found its way into nearly every aspect of our everyday lives. With this new realm of technology, we must thoroughly understand and work to address the risks in order to navigate the space and use the technology wisely. This is the field of AI ethics, specifically AI safety.

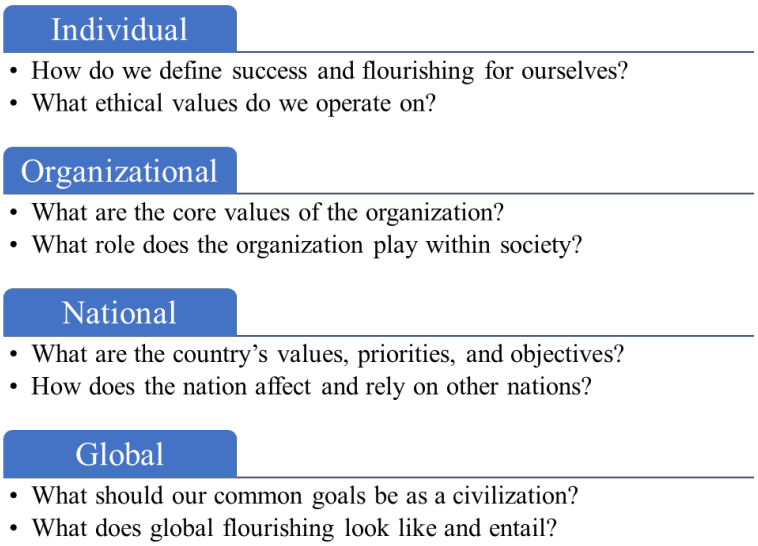
AI Alignment

AI is written to do tasks effectively and efficiently, but it does not have the abilities of judgment, inference, and understanding the way humans naturally do. This leads to the AI alignment problem: **AI alignment is the issue of how we can encode AI systems in a way that is compatible with human moral values.** The problem becomes complex when there are multiple values that we want to prioritize in a system. For example, we might want both speed and accuracy out of a system performing a morally relevant task, such as online content moderation. If these values are conflicting to any extent, then it is impossible to maximize for both. AI alignment becomes even more important when the systems operate at a scale where humans cannot feasibly evaluate every decision made to check whether it was performed in a responsible and ethical manner.

The alignment problem has two parts. The first is the technical aspect which focuses on *how* to formally encode values and principles into AI so that it does what it ought to do in a reliable manner. Cases of unintended negative side effects and reward hacking can result if this is not done properly [2]. The second part of the alignment problem is normative, which asks *what* moral values or principles, if any, we should encode in AI. To this end, we present a framework to consider the question at four levels. [3, 4]

Breaking Down the AI Alignment Problem

AI alignment is made up of value alignment problems at multiple different levels, not just in the technology itself, how it is built, and the design methods. In order for AI to truly be aligned with human moral values, all levels must be aligned with each other as well. The following is an approach to AI alignment in which the values at each level affect the others. Effects can flow downwards and upwards, and at each level, there are key questions that need to be answered.



Individual & Familial

On the individual level, the framework invites individuals and families to ask questions about values and flourishing. In our everyday actions, we are shaping our own definitions of individual flourishing—what makes life fulfilling and brings contentment. We must consider what role models and lifestyles we seek to emulate, how we define success for ourselves, what sacrifices we are willing to make, and what ethical values we prioritize.

Organizational

The organizational level refers to corporations, state and local governments, universities, churches, social movements, and various other groups in civil society. When considering alignment at this level, we must determine what values the organization operates on, what values are instilled in its products and services, and what role the organization plays within society. For institutions, important considerations are what constitutes success, what metrics are used to evaluate success, and how they are involved in the broader movements for AI alignment.

National

The next level is the national level. Each nation has either implicitly or explicitly defined values that determine the country's goals and objectives pertaining to AI. A country aiming to assert itself as a global power may invest resources into building a domestic AI industry, as well as regulate the usage of AI to moderate and nudge users' behaviors towards particular views. On the other hand, a country aiming to promote freedom may follow a decentralized approach to AI production, giving firms freedom and privacy while allowing for competition amongst firms. Alternatively, countries may try to build an AI initiative in a way that not only ensures that they are aligned with moral values, but also encourages or requires other countries to do so.

Global

Globally, humankind must think about the kind of future we want to have. The recently articulated United Nations Sustainable Development Goals (SDGs) offer a good starting point, but these goals are merely the preconditions necessary for survival and flourishing, so they are not enough [5]. A further step is needed to determine our common goals as a civilization, and more philosophically, the purpose of human existence, and how AI will fit into it. Is it to survive, raise children, live in society, seek the truth, etc.? Related to this are the end goals of economic and political structures, as well as what powerful nations and corporations need to give up in order to attend to the needs of the poor and the earth.

Putting the Levels Together

All of these levels interact with each other. Because AI typically originates from the organizational level, often in profit driven corporations, the primary motivation is often simply to make money. However, when put in the context of these other levels, further goals should become visible: 1) AI development should be aligned to individual and familial needs, 2) AI development should align with national interests, and 3) AI development should contribute to human survival and flourishing on the global level.

But other layers in the framework also interact with each other, through inputs and outputs. For example, looking at the same organizational layer from the inbound perspective, individuals can choose whether or not to buy certain kinds of technologies, nations can pass laws and regulations to control what technology companies can do, and at the global level, international pressure (for example from the UN through ideas such as the Sustainable Development Goals) can also influence technology company behavior. Of note, these levels can have intermediate levels too, such as the European Union–which is above national but below global, and which has, through GDPR, had a major influence on the internet, data, and through those, AI.

Examining the individual level, we have already seen how it influences and is influenced by the organizational level. The individual level can influence the national through elections, and the global through organizations such as the UN, although these influences are quite underdeveloped. Similarly, the

global can influence individuals through international treaties, while nations obviously exert significant control over their citizens through laws and other behavioral expectations.

Lastly, the national and global levels interact. Nations influence the global state of the Earth, for example through war and other national policies with global effects (such as energy policies which can drive or mitigate climate change.) The global level can exert power back, whether through the UN or other international expectations of national behavior.

To get a more practical view of the framework, we look at the problem of social media content moderation.

Content Moderation as an Example

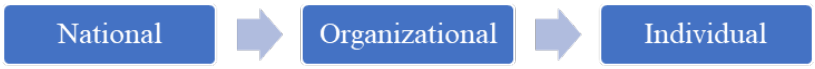
A global debate has emerged in recent years on the risks faced by internet users. User generated content is not subject to the same editorial controls as traditional media, which enables users to post content that could harm others, particularly children or vulnerable people. This includes but is not limited to content promoting terrorism, child abuse material, hate speech, sexual content, and violent or extremist content. Yet at the same time, attempts to restrict this content can seem to some like it violates user freedom of expression and freedom to hear certain kinds of expression. Organizations and governments have grappled with the feasibility and ethics of mitigating these potential harms through content moderation, while at the same time trying not to lose users who feel that their freedoms are being curtailed.

AI-assisted content moderation brings a level of speed and scale unmatched by manual moderation. A transparency report from Google (which owns the YouTube service) shows that over 90% of videos removed on YouTube between January and March 2022 were reviewed as a result of automatic flagging [6]. However, these approaches have implications for people's future uses and attitudes towards online content sharing, so it is important that the AI employed in these processes aligns with human values at multiple levels.

Using the Framework

The first issue comes from the organizational level, where there is a major misalignment between businesses and individuals. Businesses that employ content moderation (YouTube, Facebook, Google) are incentivized to maximize shareholder value, which leads to prioritizing profit over social good. For example, Facebook does this by basing its algorithm on "engagement" – the more likes, comments and shares a topic or post receives, the more it will appear on people's newsfeed. Per profile as well, Facebook can keep track of the user's behavior and habits based on engagement to feed them what they want to see. This way, users will spend more time on the site and generate more ad revenue for the business to boost shareholder value. This however leads to echo chambers and polarization, as users are not exposed to opinions that differ from theirs, ultimately affecting not only individuals and families, but

also entire nations, and even global discourse. The misalignment between organization and individuals has already proven to be dangerous with cases like Myanmar's attack on minorities illustrating the potential consequences [7].



National regulations shape how organizations moderate content, as organizations must build AI within the bounds of these regulations. A country's content moderation legislation is typically an expression of the cultural values of the majority of its citizens, which is often similar to the cultural values of its leadership, though not always. While these regulations are made by individual lawmakers and may express the values of many individual citizens, these regulations also will affect both organizations and other individuals. For example, a common good perspective might lean towards high content moderation for the sake of minimizing social harm, but at the expense of individual freedom of expression.

The question then arises regarding the alignment of cultural values with AI content moderation. We may be able to recognize where there are misalignments between national and organizational values, which in turn affects individuals. For example, in the US, where individual freedoms is a priority, there is very little content moderation regulation and it requires companies such as Facebook to only moderate things such as illegally sharing copyrighted content and criminal activity such as sharing child sexual abuse materials. Therefore, while Facebook is complying with every relevant government regulation, there have nevertheless been harmful effects on society, showing how the US government content moderation legislation is not aligned with societal needs. Cases like Myanmar also suggest that this American legislation may not be aligned with global needs, as other countries are subject to these same problems and facing the repercussions of it.

Based on the above, it might seem that the first goal for AI alignment would be to align the national and organizational levels (assuming that the organization is also aligned with individual well-being). However, this is not enough – we must also consider whether these national values are aligned on the global level, that is, whether they support global human flourishing.

The effects flow in both directions. Organizations doing content moderation sometimes respond most to individual user feedback, a powerful enough organization can have a hand in swaying national interests, and a nation or group of nations can potentially change the course of human civilization.

All in all, content moderation is a prime example of how value alignment is at work right now in society. It may not be feasible to align all four values at once, but with this framework we can identify some causes of these complex misalignments.

Conclusion

If we are to make any progress on the normative side of AI alignment, we must consider all levels – individual, organizational, national, and global, and understand how each works together, rather than only aligning one or a few of the parts. Here we have presented a framework for considering these issues.

The versatility of the framework means that it can be applied to many other topics, including but not limited to autonomous vehicles, AI-assisted clinical decision support systems, surveillance, and criminal justice tools. In these hotly contested spaces with no clear answers, by breaking them down into these four levels, we are able to see the parts at play in order to create ethical and aligned AI. Perhaps then we can sleep easy knowing we'll be safe from a Terminator in our distant future.

Works Cited

[1] "Terminator 2: Judgment Day," *Carolco Pictures*, 1991. Available at: www.hulu.com/movie/terminator-2-judgment-day

[2] Dario Amodei, et al., "Concrete Problems in AI Safety," *arXiv*, 25 July 2016. Available at: https://arxiv.org/abs/1606.06565

[3] For a brief previous presentation of this work, please see "A Framework for the AI Alignment Problem," *Student Showcase 2022*, Markkula Center for Applied Ethics, Santa Clara University, May 17, 2022, minutes 50:15-54:20. Available at: https://www.youtube.com/watch?v=UIkuVq83o48&t=3018s

[4] For another take on the problem see the "multiscale alignment" section of Max Tegmark's interview with the *80,000 Hours Podcast* where he describes a similar sounding idea that he developed. Tegmark's framework does not yet seem to be published, so we cannot know in exactly what ways his and our frameworks are similar or different. Robert Wiblin and Keiran Harris, "Max Tegmark on how a 'put-up-or-shut-up' resolution led him to work on AI and algorithmic news selection," *The 80,000 Hours Podcast*, July 1st, 2022, minutes 1:13:13-1:51:01. Available at: https://80000hours.org/podcast/episodes/max-tegmark-ai-and-algorithmic-news-selection/

[5] "THE 17 GOALS - Sustainable Development Goals - the United Nations," *United Nations*. Available at: https://sdgs.un.org/goals

[6] "YouTube Community Guidelines enforcement," *Google*. Available at: https://transparencyreport.google.com/youtube-policy/removals

[7] Paul Mozur, "A Genocide Incited on Facebook, With Posts From Myanmar's Military," The *New York Times*, 15 Oct. 2018. Available at: https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html