When AI uses signals inherent to a person to predict behavior: protecting human autonomy, self-worth and relationships

Adarsh Mital

There are numerous ethical principles that govern the decision-making process of state actors and others with authority, e.g. a business or professional. Sometimes, the concern is the impact of a decision, at other times what information was considered or how it was obtained. Some of the principles are grounded in individual rights, with corresponding duty not to infringe. Others are more Consequentialist in nature, i.e. limits or rules to ensure that the conduct is moral in the sense its effect is more favorable than unfavorable to some combination of actors or subjects. These ethical principles manifest in:

1. The Constitution. For example, the Fourth Amendment implies a right to privacy, restricting what information can be collected. The Equal Protection Clause of the Fourteenth Amendment is interpreted as prohibiting taking race into account in policies.
2. Laws. For instance, Title VII of the Civil Rights Act of 1964 prohibits employers from considering “protected attributes” of race, color, religion, sex, or national origin. The Electronic Communications Privacy Act restricts what information can be intercepted and collected. Some restrictions were loosened by the USA Patriots Act.
3. Voluntarily adopted operational guidelines. For example, not taking criminal history into account while evaluating for employment.

Technological progress changes the nature of data that can be collected as well as the decisions that can be taken. This often calls for expanding how ethical principles apply to new situations. The ensuing debate is first seen in papers and proposals from ethicists and policy makers. Next, new cases go to court where judges may extend the scope of the Constitution and Laws, though mostly we receive dissenting opinions to add to the debate.

As Artificial Intelligence has developed, the US Government has not yet passed laws to govern it.There is a lot of debate, however, about the rules and guidelines that constitute ethical AI. A few have been voluntarily adopted, mostly by companies that create AI and sometimes by companies that apply AI in areas like granting credit or screening candidates.

This ethical debate currently centers around ensuring that a state or business actor cannot use AI to achieve what they are restricted from doing themselves. A minority of commentators warn about the risk of AI degrading humanity, but because many of them also talk about Artificial Generalized Intelligence – which is seen as futuristic – people dismiss them.[[1]](#footnote-2)

This paper is about a near-term risk that few people foresaw, a technical capability that is becoming imminently possible due to advances since 2019. This risk falls in between the two current strands of discussion of ethical AI. Below, we will examine the gaps and also discuss how society may reduce negative outcomes while preserving some of the potential benefits.

# AI Ethics debate today

Much of the discussion of ethics for AI revolves around ensuring transparency, responsibility and privacy, justice and fairness, and non-maleficence in AI’s information collection and decision making[[2]](#footnote-3). Many commentators, particularly from AI creators like Google and Microsoft, discuss these goals in terms of the “technical” problems that AI practitioners should solve:

Transparency: explain how an algorithm arrives at its decision.

Responsibility and Privacy: (a) Seek informed permission from people before collecting data, and (b) prevent models from considering “protected attributes”, either directly (a person’s race) or indirectly (by co-occurring information like where they live).

Justice and Fairness: Prevent inadvertent bias in algorithms, i.e. ensuring that algorithms do not learn from data that represents today’s inequities.

Non-maleficence: Sometimes instantiated as preventing algorithmic bias and safeguarding privacy, though sometimes also as not harming the fabric of society[[3]](#footnote-4).

Legislators and politicians sometimes talk about the larger notion of fairness: how everyone gets access to the benefits of AI, and how to mitigate the potential impact of automation on labor markets. It is possible to think of this as Consequentialist and Utilitarian. There is nothing inherently forbidden to AI, as long as its consequences are more favorable than unfavorable to society at large. Society is Bentham’s party whose interest is in question. There is some discussion on the need to avoid doing harm (Non-maleficence), but it is at the level of slogans: “*The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends*.”[[4]](#footnote-5)

Many AI ethicists, too, have no issue with AI venturing into almost any area of decision making. They find themselves able to justify AI to determine hiring, guide prison sentences or evaluate medical treatment. All they ask for is that the AI be un-biased and fair (e.g. don’t amplify inequities associated with poverty inherent in today’s data, e.g.), be explainable and accountable[[5]](#footnote-6), not use protected attributes[[6]](#footnote-7) and not learn from data obtained by ignoring a right to privacy or data ownership[[7]](#footnote-8). AI leaders like Microsoft and Google have adopted this thinking and have surfaced in tools like Fairlearn open source package[[8]](#footnote-9) and Fairness Indicators[[9]](#footnote-10) to help data scientists test whether an AI model uses protected attributes as operative parameters, or whether its outcomes disproportionately impact certain groups.

However, a newer strand of AI ethics considers AI in the largest sense of its potential to pollute or disturb human thought. For example, every party in a conversation can influence the other. If AI systems have a dialog with consumers, they may influence language, perhaps human thought itself[[10]](#footnote-11). An analogy is how social media has affected culture and relationships; with AI’s ability to do more than just serve feeds and actually converse, it can have a far greater impact.

## Moral Absolutist thinking about AI ethics

Away from the mainstream, a few people have taken a more stringent stand that is closer to Moral Absolutism. For them, some decisions belong in the human realm, e.g. the ability to engage in warfare and target people[[11]](#footnote-12) must not be entrusted to AI – whether the algorithms are accurate or not, biased or not, explainable or not.

## Government stance leans towards the technical view of AI ethics

The US government and Congress have not agreed with the moral absolutists. They have not yet forbidden AI from any activities. However in response to AI companies calling for regulation[[12]](#footnote-13) from the technical view of AI ethics, the government has published draft rules[[13]](#footnote-14).

## Distinction from social policy to manage AI’s impact

There is public concern about the impact of AI on jobs. Both Pete Buttigieg and Andrew Yang talked about this during the Democratic presidential primaries. Of course, they both came on the side of wanting even more AI that is better than China’s. Buttigieg suggested that displaced workers get a universal basic income as compensation. This social policy debate is not about the morality of what AI should be allowed to learn or what kinds of decisions it should be allowed to take. Just how we can manage to live with the inevitable outcomes.

# Disruption: Brute-force mapping to predict complex phenomena from small signals

While some alarmists have been looking for signs of AGI, a brute-force and near-simplistic approach to AI has changed the character of the industry and taken people by surprise. Here is an example.

GPT3 from OpenAI has created an AI model that has learnt all frequently read text available on the web: every sentence, every paragraph, every variation of thought. Imagine a high school senior writing an essay on modern politics. The student gives GPT3 the beginning of their thought, the first three sentences. The system fills in 3 whole paragraphs. Grammatical, and to a high-schooler’s eyes, coherent and complete. More than that, people report that the sentences eerily similar to how they write, with peculiarities that are not obvious in the starter 3 sentences. It is almost as if the three sentences contained enough clues to predict the person’s train of thought.

The trick is under the covers, first revealed by Google in the BERT Large transformer architecture. GPT3 has applied it and placed an internet worth of text in 100 billion dimensions (3 orders of magnitude more than familiar convolutional neural networks.) When information is put in these many dimensions, things that are alike in any way (the author’s tendencies, style, emotion, flow, tone, pace, the subject, the actors, the locales) cling together in some dimension.

If someone really tests GPT3, they can see flaws. Give it non sequiturs, non-sensical word combinations, GPT3 will confidently produce grammatical flowing writing, but with nonsense sprinkled throughout. Yet this will certainly improve.

## What we are learning about the ability to model the human’s world

Industry is beginning to realize that with this brute-force approach to AI, at least some parts of people’s worlds, parts we associate with willfulness and volition, can be plausibly modeled:

*People can be characterized as similar to others along certain dimensions: Their actions or biology may resemble one other in situations that are similar in some dimension.*

*Examined closely, there are enough hints in what a person would consider a small signal. It does not take much to make predictions of complex chains of events.*

## A blueprint for companies to invest to map an entire aspect of the human world

We have a path here for companies and VCs who are willing to invest big. Now they can create AI that maps entire aspects of the human world. Sometimes, this means mapping externally visible behavior of a type that people think as volitional or controllable. At other times, the map is about the inner workings of the human body. The two may or may not be related, they are included here as examples of complex systems that were thought too difficult to model:

* Map the world of engineering design by learning over billions of drawings and millions of produced artifacts. Take an unfinished sketch of an equipment and generate what the engineer would put inside.
* Map the plays coaches call by learning from the geometry of every play, player and coach profiles, photos and commentary, and pre- and post-match interviews[[14]](#footnote-15). Use that to predict not just the next play, but an entire quarter, with moves and countermoves.
* Map the human biome by learning from millions of patients’ blood work and medical history. How every bacterium, every colony interacts with the host, including all the variations of genetics and diet across the human population. Use it for “precision medicine,” designing molecules based on a prediction of that body’s reaction.
* Map the 3D shapes of proteins (learning from 170,000 known shapes and 180 million unknown ones) to predict how a newly invented biomolecule would fold up inside the body in complex 3D shapes. Even though amino acid chains can theoretically fold in 10 to the power 300 shapes, predict the geometry down to the width of the smallest atom[[15]](#footnote-16).

Every creator of such AI knows that it will take years to get the predictions sufficiently accurate for widespread industry use. The key is to find areas where some people are willing to pay for partially accurate answers: often because that is a lot better than no answer at all, and the consequences of inaccuracy are small.

The risk here is that someone will take this blueprint and apply it to the world of human behavior. And that it will be good enough (predictions plausible enough) for people to adopt subsequent systems in situations when one person can judge another, and deny them opportunity, freedom or companionship.

# Mapping long-term human behavior is a natural progression

It is a small stretch to imagine that someone like Google or Softbank will fund the creation of a massive map of human behavior. Pay a hundred thousand volunteers to allow cameras in their space, wear sensors, allow medical tests, and then capture signals for extended periods regarding:

1. What they say and write, who they interact with, location and setting, facial expressions,stance, gait, pulse rate (which AI can already discern from a video stream), intonation in the voice, the body’s magnetic or FM signature.
2. Their DNA and other so-called “omics” gleaned from regular blood tests and EKG.
3. Their daily reports of what happened at college, home or work.

Indeed the market already pays for conventional AI that predicts only a few, vague events:

* Recruitment systems using handwriting to predict honesty, and social postings to predict “Potentially Unlawful Activity,” and “Potentially aggressive verbiage”[[16]](#footnote-17).
* Parole boards using AI to predict the risk of a prisoner re-offending.

The promise – or concern for an ethicist – is that this newly mapped AI will now be able to predict a far broader range of behavior: what a person may do in a particular situation, down to a whole sequence of events that may last hours, days or months.

## AI predictions inserted into everyday interactions

If an app is able to take a few signals – a Zoom capture by a high resolution camera, and their Facebook, Instagram and LinkedIn – and make interesting predictions, I can imagine individual use of it, perhaps multiple times a day, to predict the actions of others they have relationships with:

* If music gets loud, and beer has run out, my friend will pick a fight at this evening’s party.
* If my co-worker does not sleep well, she’s going to be excessively critical of my work in a group setting
* If the Lakers lose badly, my boyfriend will likely withhold affection the rest of the evening.

Take a professional setting. Until now, AI has been used for screening recruiters. Now, imagine a manager uses this new AI to choose who to assign to a new role in Beijing, one where they will have to eat unfamiliar food and live in a cramped apartment.

Where previously AI was used only by the parole board, now even the local police may want to judge a newly released prisoner on a regular basis. Given recent news, an economic downturn, cloudy weather, and the type of person who has moved next door, predict if this person is likely to re-offend and needs intervention.

# Potential impact of removing the wall between Goffman’s front- and back-stages

This new AI will be able to see through what Goffman refers to as the more ‘public’ front stage and observe the back stage even when the actor does not want you to. Goffman highlights that this ability to decouple our back-stage self from our front-stage persona is ingrained in our existence. We can surmise that it is essential to our psychological well-being, sense of self-worth, and to our relationships.

It is conceivable that some may adapt to the loss of this distinct separation between the front stage and back stage-so long as they are not the only ones forced to do so. An example is how it is now commonplace for people to stream their entire lives for others to see, revealing intimate things like breakups and not worrying about how they are being judged. This exposurecould become the new norm.

However, many others will react badly to this lack of control over how to deterministically change how others perceive them. Some don’t leave their house if they have no good clothes to wear, now they have no good clothes for their persona.

In work and prison settings, the power relationship becomes even more asymmetric. The candidate or the prisoner will feel helpless – particularly if they know they are being judged for characteristics that people believe they cannot change or self-improve.

# Analysis: how ethical principles might be applied to this new AI capability

## Coverage by today’s privacy protections

So long as the model predicts behavior using only the information people willingly transmit by getting on FaceTime, visiting an office or house, or posting a video, privacy laws will likely not protect the large variety of information that will be used by this AI.

Today, in Washington state, people have to ask permission before recording someone’s video call, but there are no restrictions on what someone may do by studying the recording. As long as the other person consents to being recorded, AI is just another way of studying the video.

It is hard to imagine that courts will prevent an employer or a university from interpreting an interview or a posted video using AI. Madden, Gilman et al[[17]](#footnote-18) refer to consumer reports that use social postings to characterize applicants’ future behavior. The employer faces no liability if reports are wrong. Courts have also determined that the “legitimate business need” provision of Title VII of Civil Rights Act justifies adding third party information to build up a richer picture of the candidate, e.g. what someone has bought in the past, where they have lived.

## Coverage through “protected attributes”

AI models can thoroughly checked for whether they take protected attributes into account, or whether they have disproportionate impact on protected groups.

It is possible a few behavioral tendencies have a high correlation to race, religion, etc. Those may be covered as indirect ways of usingprotected attributes. However, any behavioral predictions that are more general in character and less dependent on these limited attributes will not be covered under this umbrella.

It will require much debate and a lot of evidence of consistent harm to identifiable groups for society to extend the list of protected attributes.

## **A Utilitarian analysis**

Even if we cannot easily influence laws, if there is enough consensus amongst ethicists and enough public awareness, then many responsible and reputation-conscious businesses will voluntarily adopt them as policies. Typically, corporations, being economic entities, have a Utilitarian lens. In the case of AI, this means avoiding technology uses that most people consider harmful but preserving others that most think are beneficial. There are plenty of foreseeable situations where deploying AI to predict human behavior will have a positive effect. For example, recently discharged army veterans are at a higher risk of suicide than the general public. The VA is using AI to locate people at risk, so they can help them secure poorly placed guns or ensure they take their medicine[[18]](#footnote-19). There are enough false positives that these calls can be intrusive. So one can imagine how a more accurate model of human behavior can improve this care. Companies will want to preserve such uses.

At the other end of the spectrum, many tech companies have voluntary rules that they will not use facial recognition to assist in warfare or detection of possibly undocumented immigrants. The way they apply these rules is by not selling technology to certain government agencies. Once the impact on people of AI that predicts behavior becomes clear, it is possible that tech companies will adopt similar policies about selling AI for uses that may predict future conduct and on its basis deny people job or housing opportunities, or even freedom.

However, such voluntary limits by end purpose are hard to enforce. A tech company may supply AI to another company that creates the end user application, unknown to the AI supplier.

## Grounding in human autonomy and self-determination

Some commentators view thenormative principle of autonomy as a personal right and central to how society functions[[19]](#footnote-20): “Yet the risk is that AI systems may erode human self-determination, as they may lead to unplanned and unwelcome changes in human behaviours to accommodate the routines that make automation work and people’s lives easier. AI’s predictive power and relentless nudging, even if unintentional, should be at the service of human self-determination and foster societal cohesion, not undermining human dignity or human flourishing.”

A similar view is expressed in the context of how search and internet companies can create a picture of a person that the said individual did want to give: Systems that use information and “sell it” on “behavioral futures markets” expose our lives to others without our consent. “In losing decision rights, we lose privacy, as well as autonomy and self-determination[[20]](#footnote-21).”

However, as always, ethicists can apply different normative principles to the same problem. The following view from Jobin et al[[21]](#footnote-22) requires very little restraint by AI creators: ‘Freedom and autonomy are believed to be promoted through transparency and predictable AI, by not “reducing options for and knowledge of citizens”, by actively increasing people’s knowledge about AI.’

Some ethicists bridge both the above normative principles: there should indeed be a dialog between the public and the creators of AI so that consumers are knowledgeable about AI; but that might not be enough[[22]](#footnote-23). Additional ethical guidelines are needed to examine AI for its overall impact on how people interact with each other and function in society. These ethicists warn that AI will evolve faster than, and will be adopted faster than, people’s ability to assess the changes and prepare to either deal with or prevent the consequences.

The question is which of these views, i.e. which of the normative principles the views are rooted in, will carry greater weight.

As I have outlined by potential scenarios, the risk to autonomy is high enough and immediate enough that ethical guidelines governing this new AI should not be dominated by Utilitarian thinking. Instead, the first consideration should be to prohibit both the creation and the use of AI that sees through to Goffman’s back-stage, thus preserving the sense of self-determination that is integral to a fulfilling life.

## How bioethicists anchor the ban on human cloning in the principle of autonomy

Lastly, there may be hard limits to society’s tolerance for how much AI to predict behavior will be allowed to affect and change the human condition. There are conceivable situations where the predictions become oppressive enough to trigger the thought of moral repugnancy, at which time emotional viewpoints might trump deliberative reasoning.

For a parallel, consider the ethics of genetic engineering. Much of the debate about cloning and genetic engineering is conducted in the familiar language of autonomy, consent, and individual rights[[23]](#footnote-24). A few commentators have invoked Consequentialist principles: Ronald Dworkin, for one argues that there is nothing wrong with the ambition “to make the lives of future generations of human beings longer and more full of talent and hence achievement.” However, Consequentialist arguments have been overwhelmed by principles that seem rooted in moral repugnance: Jurgen Habermas[[24]](#footnote-25) worries that even favorable genetic enhancements may impair the autonomy and individuality of children by pushing them toward particular life choices, hence violating their right to choose their life plans for themselves.

The latter view has prevailed: human cloning is forbidden in every genetics lab in USA, without consideration of whether it has been proven safe in animals and is projected to be safe.

My takeaway from the human cloning debate is that there is a point when a prospect is so disturbing to how we live, or how we want to live – volition, self-determination, individuality – that people reach for the strongest absolutist arguments against it. And lawmakers and businesses are compelled to act.

# Conclusion

We are far from Artificial Generalized Intelligence. The threat of domination by self-aware machine intelligence is still not on the horizon.

However, what is much nearer is the ability to inject AI in the middle of how people evaluate each other or how they choose to reveal themselves to others.This injection can cause people to be treated by business, authorities and even friends in the form of deterministic machines, as opposed to self-willed and constantly evolving humans.

In my view, it does not matter whether AI that predicts behavior will be perfectly accurate. Even assuming eventual accuracy, the issue is the effect it will have on human relations and on the human condition itself.

I argue that the principles most able to provide adequate safeguards will be grounded in human autonomy and self-worth.

1. Rohde, K., Vukovic, R., Zeldich, M., Ramesh, S., Hershkowitz, J., & Farkas, G. (2018, June 13). Benefits & Risks of Artificial Intelligence. <https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/?cn-reloaded=1> [↑](#footnote-ref-2)
2. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399. doi:10.1038/s42256-019-0088-2 [↑](#footnote-ref-3)
3. Schrader, D. E., & Ghosh, D. (2018). Proactively Protecting Against the Singularity: Ethical Decision Making in AI. IEEE Security & Privacy, 16(3), 56-63. doi:10.1109/msp.2018.2701169 [↑](#footnote-ref-4)
4. AI Principles. (2018, April 11). Retrieved from https://futureoflife.org/ai-principles/ [↑](#footnote-ref-5)
5. Satell, G., & Abdel-Magied, Y. (2020, October 20). AI Fairness Isn't Just an Ethical Issue. <https://hbr.org/2020/10/ai-fairness-isnt-just-an-ethical-issue> [↑](#footnote-ref-6)
6. Harned, Z., & Wallach, H. (2020). Stretching Human Laws to Apply to Machines: The Dangers of a "Colorblind" Computer. Florida State Law Review, 45(617). [↑](#footnote-ref-7)
7. Holstein, K., Vaughan, J., Daume, H., Dudik, M., &amp; Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In ACM CHI Conference on Human Factors in Computing Systems. ACM. [↑](#footnote-ref-8)
8. Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., . . . Walker, K. (2020, September 22). Fairlearn: A toolkit for assessing and improving fairness in AI. <http://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai> [↑](#footnote-ref-9)
9. Xu, C., & Doshi, T. (2019, December 11). Fairness Indicators: Scalable Infrastructure for Fair ML Systems. <https://ai.googleblog.com/2019/12/fairness-indicators-scalable.html> [↑](#footnote-ref-10)
10. Schrader & Ghosh, ibid. [↑](#footnote-ref-11)
11. Freedberg, S. (2019, September 18). Campaign To Stop Killer Robots 'Unethical' &amp; 'Immoral': Bob Work. <https://breakingdefense.com/2019/08/campaign-to-stop-killer-robots-unethical-immoral-bob-work/> [↑](#footnote-ref-12)
12. Smith, B. (2020, May 05). Facial recognition technology: The need for public regulation and corporate responsibility. <https://blogs.microsoft.com/on-the-issues/2018/07/13/facial-recognition-technology-the-need-for-public-regulation-and-corporate-responsibility/> [↑](#footnote-ref-13)
13. Request for Comments on a Draft Memorandum to the Heads of Executive Departments and Agencies, "Guidance for Regulation of Artificial Intelligence Applications". (2020, January 13). <https://www.federalregister.gov/documents/2020/01/13/2020-00261/request-for-comments-on-a-draft-memorandum-to-the-heads-of-executive-departments-and-agencies> [↑](#footnote-ref-14)
14. Oved, N., Feder, A., & Reichart, R. (2020). Predicting In-Game Actions from Interviews of NBA Players. Computational Linguistics, 46(3), 667-712. doi:10.1162/coli\_a\_00383 [↑](#footnote-ref-15)
15. AlphaFold: A solution to a 50-year-old grand challenge in biology. (2020, November 30). <http://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology> [↑](#footnote-ref-16)
16. Madden, M., Levy, K., &amp; Marwick, A. (2017). Privacy, Poverty and Big Data: A Matrix of Vulnerabilities for Poor Americans. Washington University Law Review, 95(53). [↑](#footnote-ref-17)
17. Ibid. [↑](#footnote-ref-18)
18. Carey, B. (2020, November 23). Can an Algorithm Prevent Suicide? <https://www.nytimes.com/2020/11/23/health/artificial-intelligence-veterans-suicide.html> [↑](#footnote-ref-19)
19. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., . . . Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines, 28(4), 689-707. doi:10.1007/s11023-018-9482-5 [↑](#footnote-ref-20)
20. Shaw, J. (2016, December 18). The Watchers. <https://www.harvardmagazine.com/2017/01/the-watchers> [↑](#footnote-ref-21)
21. Ibid. [↑](#footnote-ref-22)
22. Schrader & Ghosh ibid. [↑](#footnote-ref-23)
23. Sandel, M. J. (2005). The Ethical Implications of Human Cloning. Perspectives in Biology and Medicine, 48(2), 241-247. doi:10.1353/pbm.2005.0063 [↑](#footnote-ref-24)
24. Habermas, J., &amp; Habermas, J. (2003). The future of human nature. Cambridge, UK: Polity. [↑](#footnote-ref-25)