A Human Values Framework for Managing AI Development

By Chip Carter

When Smartphones started becoming widespread in the late 1990's and 2000's, there was no broad awareness of the legal, social, psychological and moral threats that would arise. At the time I was a software engineering director at Lotus Notes, immersed in technology every day. Unlike my colleagues, I had a background in ethics, philosophy and cognitive studies. As I watched smartphone usage and habits unfold, I remember thinking, "Have we thought this through? Are we prepared for how this will change our mindsets, behaviors, interactions, and socio-economic structures?" In hindsight, the answer was an indisputable "NO". But like even the most thoughtful and cautious people, the practical benefits of smartphones intrinsically forced many of those questions to recede into the background of daily life.

Now, with more than two decades of smartphone use behind us, we are aware – and not just anecdotally, but scientifically aware -- of the legal, social, ethical, and economic realities and fallout of smartphone usage. Just to briefly recite negative outcomes in the psychological space, those include:

- Increased anxiety levels and depression
- Difficulty sleeping
- Isolation from social interactions
- Decreased performance at work or school
- Relationship issues

And again, I haven't touched on issues here in the social, ethical, legal, or economic space.

So why am I reciting the negative impacts of smartphone adoption and usage at an AI conference? Because I believe that the history of smartphone usage provides a recent, relatable, relevant and science-based example that can be extrapolated to Artificial Intelligence. We can use our experience and knowledge here to inform a general approach or what I would describe as a human-value-based framework for decision-making and preparation for all of the myriad implications of AI.

I'm not going to delve into the specifics of all of the specific threats and challenges that AI poses to the modern world. I'm less qualified to do that than other speakers here, and it's a very long list that would take more than 7-8 minutes. Again, my intent is to describe a high-

level set of principles – based on our history with smartphones – that could serve as a foundation for meeting these threats and challenges.

For each principle – and they all closely connected -- I'll also put forward human values that we should cultivate in order to respond effectively to the potential benefits and detriments of artificial intelligence.

Principle One: We don't know what we don't know.

We already know a lot about AI and its potential or real impacts, like job displacement, threats to privacy, security issues, accountability and explainability issues, bias, misinformation, balancing ethics with competition, and the list goes on. But we must admit that what we know probably pales beside what will actually happen, and what we don't know. Specifically, we cannot predict what *forms* these threats will take or the how deeply the possible threats will actually impact us. The 20th and 21st centuries are replete with examples of technologies that produced poisonous and unpredictable fruit with severe and widespread impacts. AI has come to life quickly, growing from an infant to a volatile teenager where narrow intelligence is ubiquitous and general intelligence is on the horizon. There are still many questions about where AI will go and how it will be used.

Human values to cultivate in response: Curiosity, openness, humility, resilience.

In the face of its unpredictability, we need to be curious and open about the future of AI. We need to admit new (credible) information about AI readily and not cling to old paradigms or try to force-fit new data into old paradigms. And we need to avoid hubris or over-confidence in our ability to predict outcomes for two reasons: 1) we're not good at it; and more importantly; 2) hubris works against real and efficient learning. We need to encourage a beginner's mind and consistently question our assumptions and expertise. Finally, agility and resilience are needed. Dwight Eisenhower once said, "Plans are useless, but planning is indispensable". We live in a VUCA world; VUCA is an acronym that stands for Volatile, Uncertain, Complex and Ambiguous. Given the VUCA nature of where we find ourselves with AI, climate change, globalization and other realities, it's important to shape both preventative and curative measures flexibly, running intelligent experiments, be open to failure, and observing and repeating what works and learning from what doesn't work.

Principle Two: What's New is Old.

Much of what we are seeing and will see in terms of fallout has precedents in recent and not-so-recent history. If we're all bent out of shape about the potential for job displacement -- for example -- we should both relax and be concerned. It has already happened many times over in the course of history. Remember the "secretarial pools"? =)

We've seen instances where technology has augmented human work or replaced human work. Smartphones and social media provide us many examples of the "what's new is old" principle. We've seen an abundance of social isolation and misinformation – both threats which AI can potentially exacerbate – in our lifetime.

Human values to cultivate in response: Respect for history and the darker potential of human nature and imperfect institutions.

In America, after Barack Obama left office, some of us were shocked that white supremacy and misogyny resurfaced with great force. We can succumb to the idea that historical progress is a straight line inclined up, where technology, social norms, and socio-economic structures march side-by-side toward a brighter future. In reality, history is cyclical and only progressive over large arcs. In every age, we see both regressions *and* progress that mirrors something in the past. We need to respect the reality of these regressions, and believing unequivocally that they can happen, frame our decisions around their potential for harm.

Principle Three: Social, legal, economic and psychological norms and structures will always lag behind technology.

It's easy to observe the reality of this in every age. Again, smartphones provide a perfect example. A related corollary to this principle is that we are generally myopic (near-sighted) and will dismiss or ignore future threats in favor of immediate benefits (like profit, convenience, or pleasure). That was certainly the case with cell and smartphone development, which provided us with entertainment and powerful practical tools to replace more laborious and flawed alternatives.

So technology just happens. We assume that it will be put to good use and are often blind to destructive ways it is eventually harnessed. Frequently, there is insufficient partnership between technologists and decision-makers on social, economic and political fronts that might check its progress or direction. And there is almost always profit-motive behind technology development, which can blind powerful people and institutions to their stated priorities. Technology therefore moves forward, generally unimpeded and more consistently progressive than its partners.

Human values to cultivate in response: Prioritization of compassion and equality, democratic decision making and transparency, blending utility with virtues and rules.

Halting or radically slowing technology development is both unrealistic and also potentially harmful in a fast-evolving world threatened by realities like climate change that could be positively mitigated by new technologies. But many of us live under the illusion or

assumption that beneficial utilization follows technology development. To prevent damage from this historical fiction, we need to prioritize compassion and equality over unbridled technology development or profit. This means fully embracing the idea the every human-being is equally important and should not be harmed from technology or profit-related decisions. Global citizenship is more important than national citizenship. What benefits one group and harms another is a lose-lose, not a win-lose. We need more broadly representative decision-making structures that either do not exist or are operating badly to broaden their perspective and involve both technology and non-technology partners. Closing the gap between technology and healthy social advancement means *real partnerships* where technologists and non-technologists educate each other transparently and work iteratively to frame problems and craft solutions together. Finally, new ethical frameworks must be created and adopted which balance utility, the cultivation and preservation of human virtue, and inviolable rules which protect human health, dignity and equality. These frameworks exist in some places, and work well. They can be used as the foundation for better decision-making models, helping to avoid cynical conclusions and implicit assumptions like "all progress creates collateral damage". We are already witnessing that kind of cynicism and/or myopia with respect to AI development. For example, Satya Nadella has recently argued in favor of brisk AI development in the face of job/worker displacement, arguing that AI can democratize access to new skills, helping displaced workers find new jobs. At best this point of view is incomplete and myopic. At worst it denies real threats and cynically prioritizes and assumes that certain technology-driven outcomes are inevitable.

Principle Four: It should be less about technology and more about do-no-harm goals.

This principle is really an inevitable conclusion from the preceding principles (if you are still awake and believe them). I'm fully aware that I've painted with broad strokes and have probably generated more questions or doubt than concrete solutions. Finally, I'm aware of the notable scarcity of AI impact specifics in this paper. But this is because it's really not about the technology outcomes we can predict or – on the other hand – can't even now imagine. It is precisely the failure to be deeply intentional about purpose before action or the failure to be completely clear about real values, goals and priorities that has plagued technology development efforts in the past. Or more precisely, plagued societies and groups of people who have had to live with harmful consequences of unchecked technology development. As a technologist for over 30 years, I certainly advocate for technology. But what is truly naïve and short-sighted and unrealistic is to move forward without a value-based framework which might prevent unnecessary harm to us and the

rest of the planet. And might badly compromise the profound possibilities we envision with Artificial Intelligence.