Daron Acemoglu

A post about Pope Leo XIV's encyclical on AI. Why the Pope is right, but perhaps not right enough. Artificial intelligence is reshaping the world in front of our eyes: how we communicate, how we access information, how we work, how income and status are distributed among us, and soon how we fight and kill each other. Yet the public conversation about AI remains stuck on the minutiae of competition between labs, or on a false dichotomy between AI as a “stochastic parrot” with no real capabilities and AI as an alien superintelligence poised to take command of humanity. The more important questions are about what we want from AI, and whether our current mindset, institutions, and control mechanisms are equal to the task of steering it toward our welfare. It is refreshing, then, that a bold and powerful voice has weighed into this debate: Pope Leo XIV. As an economist who has long argued that technology is a matter of choice rather than fate, I find Leo’s intervention welcome and, on most points, on target. But on the most consequential question of what AI should actually be designed to do, Leo stops short. Secular readers may bristle at the encyclical’s opening invocation of the Tower of Babel. They would be mistaken to stop reading there. Leo goes much further than most pundits, journalists and policymakers in the United States by recognizing that what happens to AI, and hence to humanity, is a under our control. There are multiple possible paths for AI, and which one we take will have sweeping consequences. He is also ahead of many commentators when he writes forcefully and unequivocally that “technology is never neutral, because it takes on the characteristics of those who devise, finance, regulate, and use it.” These were the central themes of the book I wrote with Simon Johnson, Power and Progress: Our Thousand-Year Struggle over Technology and Prosperity. It is heartening to hear them taken up by a voice with Leo's reach. The Pope is also right to question the current trajectory of AI in warfare and law enforcement. What was taboo only a few years ago – AI-driven mass surveillance, algorithms selecting targets for killing – has become routine. Many in Silicon Valley are now calling openly for a new military-algorithmic complex centered on AI as an instrument of American hard power. Leo captures something deep and too often ignored: “Any technology that facilitates attacks without seeing the face of human beings lowers the moral threshold of conflict.” His call for the “disarmament of AI” follows directly from these observations. As he explains, disarming AI means “freeing it from the mentality of ‘armed’ competition, which today is not limited simply to the military context, but is also an economic and cognitive phenomenon.” His moral clarity in stating that “there is no algorithm that can make war morally acceptable” should be a warning to technologists rushing to design new weapons of mass destruction. Underneath these specific concerns lies a more fundamental claim: that what is technically feasible is not the same as what is good for humanity, and that the difference depends on who controls the technology and what ideology and interests guide them. Leo edges toward what I take to be the most important point about AI's future when he observes that “while AI promises to boost productivity by taking over mundane tasks, it frequently forces workers to adapt to the speed and demands of machines,

rather than designing machines to work with those who work." But here he does not go far enough. He stops short of questioning the prevailing design philosophy of AI itself: a philosophy centered on mimicking human capabilities and automating human tasks, with the ultimate goal of artificial general intelligence (AGI) that can do everything a person can. This philosophy rests on a mistake. It assumes that artificial intelligence and humanintelligence are fundamentally similar, and therefore machines should naturally take over whatever humans currently do. Yet these intelligences are fundamentally different. Humans are "one-shot" learners. We form hypotheses from a few examples, mentally simulate possibilities, and refine our understanding through a social process of trial and error. This is how children learn language - imitating a few words, generalizing, and adjusting based on how others respond. We are not, however, very good at absorbing massive volumes of information or sifting through unstructured data for relevant patterns. AI models are almost the opposite. They thrive on enormous training sets and excel at pattern recognition at scale. But they have, as yet, no genuine creativity, no real-world embodiment, and no capacity for trial-and-error learning grounded in interaction with the physical and social world. When two things are different – you shouldn't, and typically you couldn't – use one to mimic the other. If you did, you would end up with suboptimal, disappointing results. It would have been a colossal mistake, and the Chicago Bulls's legendary coach Phil Jackson would have gone down in the annals of basketball as one of the worst coaches in history, if he decided in the 1990s that because Michael Jordan was the better player, Jordan should mimic everything that Scottie Pippen and Dennis Rodman were doing in the team. The team went from championship to championship because these players worked together and complemented each other. The same applies to AI and human skills. The more productive path is complementarity – using AI to do what humans cannot, so that humans can do what they do best. An electrician aided by AI diagnostics, a nurse supported by AI in interpreting symptoms, a teacher using AI to personalize instruction for each student; these are the contours of a different AI future, one that raises rather than displaces human capability. Optimists and industry insiders will respond that automation-first AI can still benefit everyone, provided redistributive policy keeps pace. But this argument has a poor track record. Forty years of digital automation have already concentrated gains at the top, hollowed out middle-skill work, and produced disappointing aggregate productivity growth. There is little reason to expect that an even more powerful round of automation, deployed by even more concentrated firms, will end differently. We can and must demand a different design. The global stakes from the future of AI are even larger than those we can see around us in the United States. For the developing world, where billions still depend on the prospect of decent jobs as a path out of poverty, an automation-centric AI agenda is not merely suboptimal. It is simply transferring to foreclose the most important route to broad-based prosperity. The biggest failing of today's AI industry is its refusal to recognize any of this. It is guided instead by an ideology of control (the industry's own over humanity) and by a conviction that machines are uniformly better than humans. As Leo rightly notes, this failure is enabled by the fact that a handful of companies now command the future of AI. What we need is a

combination of moral clarity and a serious, society-wide debate about what AI can do and what we want it to do. That debate must move beyond exhortation toward concrete choices: antitrust action against the dominant platforms, public investment in human-complementary AI, regulation of surveillance and autonomous weapons, and meaningful rights for workers and citizens over the data on which these systems are built. The Pope's intervention makes such a debate a little more likely today than it was before. It is now up to the rest of us to carry it further than he was willing to go.