Nicolas de Condorcet and the First Intelligence Explosion Hypothesis

Mahendra Prasad

■ *The intelligence explosion hypothesis (for example, a technological singularity) is roughly the hypothesis that accelerating knowledge and technological growth radically changes humanity. While 20th-century figures are commonly credited as the first discoverers of the hypothesis, I assert that Nicolas de Condorcet, the 18th-century mathematician, is the earliest to (1) mathematically model an intelligence explosion, and (2) present an accelerating historical worldview, and (3) make intelligence explosion predictions that were restated centuries later. Condorcet provides insights on how ontology and social choice can help resolve value alignment.*

The intelligence explosion hypothesis (IEH),[1] the possibility that accelerating knowledge and technology radically changes humanity, has been continuously examined since the 1980s, although some intimations of it came earlier. In the 1950s, John von Neumann was quoted as claiming that accelerating technological progress was approaching a mathematical singularity beyond which human affairs would radically change (Ulam 1958; Kurzweil 2005). In 1965, Irving John Good contended that once a smarter-than-human AI developed, such AI could build a second-generation AI smarter than the first generation and that

this process could quickly, recursively, and indefinitely lead to an "intelligence explosion" (Good 1966; Kurzweil 2005). In 1981, Stanislaw Lem published a novel, *Golem XIV*, about a military AI that increases its intelligence, rapidly approaching a singularity (Lem 1981). Two years later, Vernor Vinge contended that rapidly self-improving AI would quickly approach a technological singularity beyond which reality is unpredictable (Vinge 1983; Kurzweil 2005). In 1985, R. J. Solomanoff described six AI milestones he believed would lead to AI asymptotically approaching infinite intelligence in finite time (Solomonoff 1985). In 1988, Hans Moravec contended that in roughly 50 years, due to Moore's law and similar trends, robotic reasoning and behavior would exceed that of humans (Moravec 1988; Kurzweil 2005). In 1990, examining evidence outside robotics, Ray Kurzweil came to similar conclusions (Kurzweil 1990, 2005).

These figures all made important, independent contributions to current IEH discourse. But Nicolas de Condorcet, the 18th-century French mathematican and philosopher, (1) mathematically modeled IEH, (2) used history to demonstrate accelerating progress, and (3) made IEH predictions eerily similar to current IEH predictions. In this article, I argue that reexamining Condorcet offers useful insights. To set the stage for that discussion, I first provide a minibiography of him.

Championed in mathematics by Jean-Baptiste d'Alembert, Nicolas de Condorcet (1743–1794) made important contributions to calculus early in his career. In the 1770s, he gradually shifted from mathematics toward political economy and French politics. In 1785, he published his *Essay on the Application of Probability Theory to Plurality Decision-Making*, arguably the first major work in the fields of multiagent systems, social choice, collective intelligence, and epistemic democracy theory.[2] Condorcet was a prominent democracy advocate and French revolutionary. Like the prerevolution Maximilien Robespierre, Condorcet opposed the death penalty and thus argued that Louis XVI should be imprisoned, not executed. His opposition to Louis XVI's execution was a major reason that a warrant was issued for Condorcet's arrest during the Reign of Terror. Condorcet went into hiding, and wrote *Sketch for a Historical Picture of the Progress of the Human Mind*, a very rough draft of a planned larger work, which argued that knowledge acquisition, technological development, and human moral progress were accelerating. Condorcet explained that he was confining himself to rough outlines in this nontechnical work and that "the [larger] work itself will offer further developments and proofs" (Lukes and Urbinati 2012, 8). Unfortunately, Condorcet was captured and died under mysterious prison circumstances in 1794, and the *Sketch* was published postumously without the proofs he'd intended for the larger work. Thomas Malthus's *Essay on the Principle of Population* (1798) was written as a response to Condorcet's *Sketch* and William Godwin's writings.

Condorcet's Singularity

Cordorcet's *Essay on the Application of Probability Theory* discussed different situations in which agents with potentially different probabilities of correctness could aggregate their information to produce better system judgments. One specific situation and result is of particular note: Condorcet's jury theorem. Roughly, the theorem is this: Suppose there is a multiagent system (with n agents) confronted with a statement that is in exactly one of two states: true or false. Suppose each agent has an independent and indentically distributed probability $p > ½$ of correctly determining the statement's state. If each agent honestly reports her determination, then the majority has a greater probability of correctness than the minority. As n increases, the probability that the majority is correct quickly and asyptotically approaches 1. This asymptote may be called *Condorcet's singularity*.

Condorcet understood that this result was only one corner case of several possible situations. However, it was the basis of his democratic theory and IEH. If the agents were humans, then if humans became more honest, more independent thinking (that is, statistically independent in judgments), more knowledgeable (that is, higher p), and more numerous, then majority rule (both officially in political elections and within scientific councils, but also unofficially in public opinion and everyday groups) could asymptotically approach perfect knowledge (that is, 1).

Thus, if background social and political institutions were put in place to promote honesty, independent thinking, and knowledge in as many people as possible, then majority rule accelerates knowledge and technological progress. This is why Condorcet argues that everyone has a right to honesty and that the only time deception is acceptable is if it likely minimizes deceptions and errors in the long run (Lukes and Urbinati 2012). This is why, in an era when long-distance communication and coordination were difficult, Condorcet asserted that general populations should vote by mail, rather than in large meetings, to minimize the possibility of voters forming independent-thought-destroying factions (McLean and Hewitt 1994). This is why Condorcet supported universally accessible instruction for all humans to increase knowledge (McLean and Hewitt 1994). This is why Condorcet advocated for universal suffrage and rights, regardless of race, class, sex, or sexual orientation, increasing the number of agents (Baker 1976; McLean and Hewitt 1994).

Condorcet's singularity is supposed to be a virtuous cycle. Small increases in knowledge (that is, p) or

population (that is, n) could lead to significant increases in the probability of group correctness; this improved group correctness would, in turn, lead to (1) improved instruction and increased p and (2) better technology to improve health and increase population (Baker 1976). Condorcet recognized that resource limitations could slow or halt population growth, but IEH could be sustained by increasing knowledge (Baker 2004; Landes 2016; Lukes and Urbinati 2012; Williams 2004).

Condorcet's Accelerating Historical Worldview

Condorcet's *Sketch* has ten chapters, which he called *epochs*. The first nine cover past and then-present human history. The tenth epoch predicts the future. The *Sketch*'s purpose is twofold. First, metaphorically, the book can be interpreted as a scatterplot with the general trend of progress accelerating. As Condorcet put it: "The progress of the sciences ensures the progress of the art of education which in turn advances that of the sciences" (Lukes and Urbinati 2012, 142). By way of example, Condorcet pointed out that typical 1790s persons finishing school could know more math than Isaac Newton (1643–1727) and that the same phenomenon could repeat for any generation of students and its respective recent previous generations' geniuses (Lukes and Urbinati 2012).

Second, the book shows how societies with relatively increased honesty, independent thinking, and knowledge in the greatest number of people positively correlate with great progress, while societies that fail in those respects don't. For example, Condorcet argued that the main guarantee for the speed and extent of ancient Athenian progress was independent thinking (Lukes and Urbinati 2012, 29) was independent thinking (Lukes and Urbinati 2012). For Condorcet, it was no coincidence that democratic societies that promoted honesty, independent thinking, and knowledge (relative to their historical contexts) were the ones in which the most progress occurred (for example, Athenian democracy, the Roman republic, Renaissance Italian city-states), while late antiquity feudal societies experienced reduced and negative progress.

Condorcet's Predictions

Two common predictions of IEHs are that humans will have indefinitely extended lifespans and that intelligence can be indefinitely augmented. Kurzweil, perhaps the foremost IEH theorist, independently made these predictions (Kurzweil 2005). But Condorcet made these predictions also. Toward the end of the *Sketch*, Condorcet articulates:

> How much greater would be the certainty, how much more vast the scheme of our hopes if … these natural [human] faculties themselves and this [human body] organisation could also be improved?.... The improvement of medical practice … will become more efficacious with the progress of reason…. [W]e are bound to believe that the average length of human life will forever increase... (Lukes and Urbinati 2012, 145–46)

With respect to intelligence augmentation, Condorcet says:

> May we not extend [our] hopes [of perfectibility] to the intellectual and moral faculties?.... Is it not probable that education, in perfecting these qualities, will at the same time influence, modify, and perfect the [physical] organisation? (Lukes and Urbinati 2012, 146–47)

Condorcet, Lost and Found

It is common for Condorcet to have discovered something and for it be rediscovered later. Condorcet was the first to give a coherent mathematical explanation of insurance (McLean and Hewitt 1994). He discovered Downs paradox, later rediscovered by Anthony Downs (Lukes and Urbinati 2012; Downs 1957). Condorcet's social choice was reintroduced to the world by Black reading Condorcet (Black 1958). When social choice was revived in the 1950s, it was focused on preference aggregation for decades before widening to include grade-based aggregation (Brams and Fishburn 1983; Balinski and Laraki 2010), while, by contrast, Condorcet shifted from preference aggregation to grade-based aggregation in less than 10 years (McLean and Hewitt 1994). In the 1980s, multiagent systems replaced single agents as the computing paradigm in AI (Alonso 1998), but Condorcet also used multiagent systems. Epistemic democracy theory has been thriving since the 1980s (Cohen 1986), but Condorcet wrote several works on it. Finally, Condorcet advanced arguments for IEH that were revived in the 20th century..

The reason that Condorcet's discoveries have been so often rediscovered much later is that most of his writings lack English translations, and his collected writings often did not include many of his extant mathematical works. The works of Condorcet that exist in full translation, like his *Sketch*, tend to be nontechnical. Groundbreaking formal modeling works, like his *Essay on the Application of Probability Theory*, still lack complete English translations. It is for reasons such as these, across many writers, that reexamining old works is scientifically significant. Claude Shannon's circuit design was inspired by George Boole's early 19th-century work. Benoit Mandelbrot noted that Abraham Robinson's late 20th-century discovery of nonstandard analysis was deeply influenced by reexamining Leibniz's works (Dauben 1995).

Condorcet has insights for us, but getting to those insights can be difficult. Like Leibniz, he has no complete set of collected works translated to English.

Machine translation researchers might find mutually beneficial synergies in facilitating the complete collection and translation of Condorcet's and Leibniz's works with humanities scholars. In addition to those points already discussed, Condorcet also offers useful insights on value alignment, the problem of aligning AI behavior with human values. Currently in science, the implicit ontology is roughly logical empiricism, where mathematical and empirical statements have truth values, but moral claims don't. This ontology makes value alignment difficult. Condorcet employed a different ontology. He asserted that through science, we could assign probabilities of truth to mathematical, empirical, and moral claims. Very roughly, Condorcet can be interpreted as arguing that a scientist could individually verify mathematical theorems with very high probabilities, physical laws with high probabilities, and moral claims with low probabilities, because it would be extremely difficult to scientifically and individually verify moral claims. However, even though moral claims might have low probabilities from the perspective of individual agents, by aggregating information across multiple agents under appropriate background conditions, we could assert some moral claims with relatively high probabilities. One doesn't have to believe Condorcet's ontology to make use of it as a pragmatic means for resolving the issue of value alignment.[3]

In these and other matters, despite hailing from over two centuries ago, Condorcet's work continues to be relevant and it continues to offer us fresh insights today.

Notes

1. IEH has several names, the most popular being *intelligence explosion* or *technological singularity*. We aren't concerned with particular IEH versions, but the family of versions. Unless noted, building on Nick Bostrom's convention, I use *IEH* to refer to this family of versions. For IEH taxonomies, see E. S. Yudkowsky (2007) and Bostrom (2014).

2. Aristotle's and Jean-Jacques Rousseau's work, preceding Condorcet, alluded to the wisdom of crowds. However, Condorcet was the first to technically demonstrate how individuals' information could be aggregated to construct higher probability collective information.

3. Using Condorcet's ontology for value alignment is beyond the scope of this article. The matter is more fully discussed in Prasad (2018).

References

Alonso, E. 1998. From Artificial Intelligence to Multi-Agent Systems: Some Historical and Computational Remarks. *Artificial Intelligence Review* 21(1): 3–24.

Baker, K. M., ed. 1976. *Condorcet: Selected Writings*. Indianapolis, IN: Bobbs-Merrill.

Baker, K. M., trans. 2004. Sketch for a Historical Picture of the Progress of the Human Mind: Tenth Epoch. By Cordorcet. *Daedalus* 133(3): 65–82.

Balinski, M., and Laraki, R. 2010. *Majority Judgment.* Cambridge, MA: The MIT Press.

Black, D. 1958. *The Theory of Committees and Elections*. Cambridge, UK: Cambridge University Press.

Bostrom, N. 2014. *Superintelligence*. Oxford, UK: Oxford University Press.

Brams, S. J., and Fishburn, P. C. 1983. *Approval Voting*. Berlin: Springer.

Cohen, Joshua. 1986. An Epistemic Conception of Democracy. *Ethics* 97(1): 26–38. doi.org/10.1086/292815.

Dauben, J. W. 1995. *Abraham Robinson: The Creation of Non-Standard Analysis*. Princeton: Princeton University Press.

Downs, A. 1957. *An Economic Theory of Democracy.* New York: Harper and Row.

Good, I. J. 1966. Speculations Concerning the First Ultraintelligent Machine. *Advances in Computers* 6: 31–88. doi.org/10.1016/S0065-2458(08)60418-0.

Kurzweil, R. 1990. *The Age of Intelligent Machines*. Cambridge, MA: The MIT Press.

Kurzweil, R. 2005. *The Singularity Is Near.* New York: Penguin.

Landes, J. 2016. The History of Feminism: Marie-Jean-Antoine-Nicolas de Caritat, Marquis de Condorcet. *Stanford Encyclopedia of Philosophy,* January 20, 2016. plato.stanford.edu/archives/spr2016/entries/histfem-condorcet.

Lem, S. 1981. *Golem XIV.* Krakow: Wydawnictwo Literackie.

Lukes, S., and Urbinati, N., eds. 2012. *Condorcet: Political Writings*. Cambridge Texts in the History of Political Thought. Cambridge, UK: Cambridge University Press. doi.org/10.1017/CBO9781139108119.

McLean, I., and Hewitt, F. 1994. *Condorcet: Foundations of Social Choice and Political Theory.* Aldershot, UK: Edward Elgar Publishing Limited.

Moravec, H. P. 1988. *Mind Children.* Cambridge, MA: Harvard University Press.

Prasad, M. 2018. Social Choice and the Value Alignment Problem. In *Artificial Intelligence Safety and Security,* edited by R. V. Yampolskiy, 291–314. Boca Raton, FL: Taylor and Francis.

Solomonoff, R. J. 1985. The Time Scale of Artificial Intelligence. *Human Systems Management* 5(2): 149–53.

Ulam, S. 1958. Tribute to John von Neumann. *Bulletin of the American Mathematical Society*, 1–49. doi.org/10.1090/S0002-9904-1958-10189-5.

Vinge, V. 1983. First Word. *Omni Magazine (*January): 10.

Williams, D. 2004. *Condorcet and Modernity.* Cambridge, UK: Cambridge University Press. doi.org/10.1017/CBO9780511490798.

Yudkowsky, E. S. 2007. Three Major Singularity Schools. Machine Intelligence Research Institute blog, September 30, 2007. intelligence.org/2007/09/30/three-major-singularity-schools/.

Mahendra Prasad is a PhD candidate at the University of California, Berkeley. His research focuses on AI value alignment, democratic theory, knowledge representation, normative social choice, and algorithmic decision theory. He contributed a chapter on social choice and value alignment to *Artificial Intelligence Safety and Security* (2018), the first textbook on AI safety. He won best graduate student paper at the fourth annual conference of the NYU Alexander Hamilton Center for Political Economy.