Maya Indira Ganesh

The Trentino Brief

The Trentino Group

In spring 2023, a group of expert scientists, educationists, academics, policymakers, investors, and technologists gathered in the town of Trentino in Südtirol, a charming Alpine region constituted by Italy, Austria, Germany, and Switzerland. They came together as a fact-finding and policy research group referred to as the Trentino Group, after the series of reports, collectively referred to as the Trentino Brief, they would collaboratively author. Their mission, funded in part by discreet European and North American family foundations, would unfold over several years to address convergent problems facing the future of AI. With AI technologies showing promise in various scientific domains, the group conäsidered harms to human populations present and future, environmental costs, disruption of work, education, and other industries to be serious matters that had to be reckoned with. This essay is pieced together by a research and secretarial assistance network adjacent to the inner circle of the Trentino Group. Not all members of this network know each other; they assembled this work collectively and anonymously based on documents secreted into a shared online folder. The essay reconstructs some of the group's early discussions for the first report of *The Trentino Brief*, which were, unfortunately, lost in the intervening years; it focuses on a set of interconnected problems, and significant opportunities, between higher education, particularly, the humanities, and generative AI.

The Collingridge Dilemma

The Trentino Group begins by drawing attention to the Collingridge dilemma, a concept that addresses a challenge in the development and regulation of new technologies.[1] It is about the timing of regulating new technology: if you try to do it too early, you might not understand the technology enough to make good rules; but if you wait until you have all the information, it might be too late to make effective changes. It is a bit like walking a tightrope, trying to balance between understanding a new technology and being able to control its impact on society. The Collingridge dilemma presents two problems associated with prediction and control of new technologies, such as, for instance, a flying car. The problem of prediction is that when the flying car is just a concept or in early development, it is hard to predict how it will impact society. Will it be safe? Will it be environmentally friendly? How will it change the way cities are built? There are just too many unknowns to predict. The problem of control is when the flying car has been around for a while and is widely used. If you discover that it presents problems such as being very bad for the environment, it is now much harder to change or regulate it. That is because it is already deeply integrated into society—people rely on it, there are businesses built around it, and there is a whole infrastructure supporting it. Changing all of that is like trying to turn a ship around; it takes time and effort. Eventually, David Collingridge urged greater

1 The definition of the Collingridge Dilemma was written with material generated by ChatGPT-4 and cross-checked with Evgeny Morozov's article of the same name; see Evgeny Morozov, "The Collingridge Dilemma," *Edge*, accessed February 2, 2024, https://www.edge.org/response-detail/10898.

consideration to controlling the social life of technology because risks and challenges will be harder to mitigate if ignored early on. The Trentino Group believes that it must negotiate a way out of the horns of this dilemma for the good of society.

Resuscitating the Dying Humanities

Early on, The Trentino Group discusses the implications of new large language models (LLMs) for the so-called decline in humanities education.[2] The group focuses specifically on the US and the UK and seizes upon a novel solution: once its technical, social, and infrastructural issues have been worked out, AI is the tide that will make all the boats rise. Could this interregnum between the AIs of now and the AIs of the future be a moment that benefits struggling humanities fields, and could the humanities be reshaped to meet the requirements of future decades?

First though, what exactly does the group mean by the decline in the humanities? A variety of historic and contemporary factors, such as shrinking labor markets, precarity and austerity brought on by the 2008 financial crisis, competition, and generally weakened economies, are affecting several aspects of contemporary life. US American anxieties about the humanities relate to humanities courses not being a mandatory requirement for undergraduates any longer; coupled with funding cuts at the university level that have reduced the numbers of tenured positions; and a perceived lack of economic return from humanities education as compared to STEM (science, technology, engineering, mathematics) fields. There are concerns that people with a humanities education are not likely to find stable employment. These concerns are investigated by the American Academy of Arts and Sciences which finds that people with advanced degrees in the humanities are more likely to find employment than those with terminal undergraduate degrees.[3] It identifies gender as a significant variable in determining employment across the humanities and some STEM fields like engineering; the widest gap (1%) in employment was between men and women with terminal undergraduate degrees in engineering; the gap between men and women with terminal undergraduate degrees in humanities education, however, was not so consistently wide; it varied across the life span. The fact of significant student debt associated with higher education in the United States cannot be ignored as a factor influencing educational and professional choices.

The situation in the UK is different. Educationist Zoe Hope Bulaitis finds there is a perceptible shift in the valuation of the humanities toward the language of revenue and consumerism: "attempts at economic justification draw attention to a lack of acceptable languages with which to publicly articulate the work of the humanities beyond financial description."[4] Hence, the UK government proposed cuts in funding for arts courses in the academic year 2021–22

2 They take the humanities to include chiefly English, languages, the literary arts, visual and performance arts, history, cultural studies, and philosophy.

3 See "The Employment Status of Humanities Mayors," American Academy of Arts & Sciences, accessed February 2, 2024, https://www.amacad.org/humanities-indicators/workforce/employment-status-humanities-majors#32114.

4 Zoe Hope Bulaitis, *Value and the Humanities: The Neoliberal University and Our Victorian Inheritance*, Palgrave Studies in Literature, Culture and Economics, (Cham, CH: Palgrave Macmillan, 2020), 13.

from £36 million to £19 million.[5] However, a 2023 report by the UK's independent think tank the Higher Education Policy Institute (HEPI) found that the humanities were in good shape with student enrolments still high.[6] "In 2020, UK humanities research activity was 49% higher than the global average [,] outperfor[ming] all other disciplinary research areas in the UK."[7] The UK also has 19 universities in the global top 100 in the *Times Higher Education* 2023 rankings for arts and humanities, including four in the top 10,[8] and 18 in the top 100 in the 2022 QS World Rankings.[9] In other words: the humanities in the UK are doing all right.

Degrees in the life sciences, statistics, or civil engineering might open fairly obvious paths to careers in those fields. However, in the US and the UK, the humanities do not just result in one singular career choice but enable people to become skilled and contribute to other professions and careers.[10] In other words, expectations of how people progress from university to jobs in the professions or sciences cannot be ported over to the humanities so neatly. Humanities training may not pay back most quickly in the workforce, but it is likely to offer resilience and longevity for longer term prospects by enabling students to move laterally across sectors. The HEPI report finds that eight of the 10 fastest growing sectors employ more arts, humanities and social sciences graduates than graduates of other disciplines; and only 14% of employers say specific degree subjects are a selection criterion.[11] For most employers, it is the level of education that is important, not the discipline. In other words, obtaining an advanced, i.e., a postgraduate degree makes for greater chances for longer-term job prospects. Moreover, as human health and wellbeing generally increase and people live longer (particularly in the Global North) there are increased requirements for employment. Even if enrolments in the humanities in these two countries are dropping, people are returning to humanities subjects later in life through lifelong and continuing education programs.

Learning with Education Bots

The *Trentino Brief*'s gambit is that there could be a new revenue stream for the humanities in making generative AI meet the social, individual, professional and collective needs of humans because of the unique kinds of training that people in the humanities get. At present, while AI is still developing, it is significantly error-prone and resource-intensive. Transformer technology in large language models like Bard, Claude, and ChatGPT works through two interrelated processes: by *attending* to chunks of content in training datasets called tokens, and by *predicting* the next

5 See "Office for Students Consults on 49% Cuts to HE Arts Courses Funding," The Cultural Learning Alliance, May 5, 2021, https://www.culturallearningalliance.org.uk/office-for-students-consults-on-49-cuts-to-he-arts-courses-funding/.

6 See Marion Thain, *The Humanities in the UK Today: What's Going On?*, HEPI Report 159 (Oxford: Higher Education Policy Institute, 2023), https://www.hepi.ac.uk/wp-content/uploads/2023/03/The-Humanities-in-the-UK-Today-Whats-Going-On.pdf.

7 Thain, *The Humanities in the UK Today*, 7.

8 See "World University Rankings 2023," *Times Higher Education*, accessed February 2, 2024, https://www.timeshighereducation.com/world-university-rankings/2023/world-ranking.

9 See Thain, *The Humanities in the UK Today*, 7.

10 See Nathan Heller, "The End of the English Major," *The New Yorker*, February 27, 2023, https://www.newyorker.com/magazine/2023/03/06/the-end-of-the-english-major.

11 See Thain, *The Humanities in the UK Today*, 9.

chunk most likely to come after it based on the analysis of the dataset. Image models are constituted by algorithms that identify how images break down in response to data stimuli called noise; this allows algorithms to discern the internal "shape" of an image. And all models rely on vast amounts of data. All the world's literature, music, art, performance, histories, and languages can be integrated into the transformer technologies within large image and language models. This would lead to unprecedented access to knowledge. There could be new markets for designing custom education bots programmed to integrate knowledge across various humanities disciplines, offering students a holistic understanding of human culture and thought, previously difficult to achieve. (Obviously, high-speed access to the internet would be essential. And, obviously, the broken and extractive copyright system would need to be overhauled to ensure equity for creative practitioners.)

The *Brief*'s authors see a role for humanities professionals in improving the quality of education bots through a higher-level secretarial guidance and assistance, correcting errors. Humans will be required, especially at first, to "tune" the delivery of AI-generated humanities material and verify both the content's accuracy and its reception. A student could receive a tailor-made education by education bots that identify personal learning styles and adapting content to suit it. This tailoring would require scores of well-trained translators, performers, voice actors, illustrators, designers, and filmmakers to transform the sum of human knowledge to the next generation(s). Of course, eventually there would be an increasingly narrow set of humans with humanities backgrounds required for this as their knowledge becomes codified and integrated into AI. And, as humans continue to generate more humanities-informed poetry, music, art, literature, history, and studies of culture, eventually, the *Trentino Brief*'s authors argue, there would need to be some arrangements made for what they refer to as, somewhat dramatically, "The Last Literature Professor." (But more on that later.)

The Trentino Brief acknowledges a challenge: learning is best done socially, among and with other humans—either teachers or fellow students, ideally both. Educationists find that social interaction is a key part of learning; if learning was just about the delivery and reception of information, then getting information from the internet would be enough. Learning also requires interactive, dialogic, and group processes. The Covid-19 pandemic lockdowns were a moment for educationists to evaluate the different factors that influence distance learning. One cross-country study[12] of early university students in Hungary, South Africa, and Wales reported that distance learning was significantly affected by a variety of factors: the kinds of technology students had (laptop versus mobile phone); how strong their internet connections were; the home or personal environment in which they learned; face-to-face and eye contact versus having cameras off (which was related to strength of internet connection); how engaging the online material was in retaining students' attention. Culture, environment, and national contexts of the pandemic lockdown influenced learning too. In other words, online learning is about more than just the delivery of information via the internet.

12 See Desirée J. Cranfield et al., "Higher Education Students' Perceptions of Online Learning during COVID-19: A Comparative Study," *Education Sciences* 11, no. 8 (2021), https://doi.org/10.3390/educsci11080403.

Satya Nitta worked at IBM and spent five years trying to build a "personal tutor" with Watson, and is convinced that, even with newer generative AI, such an approach will not yield the intended results: "We'll have flying cars before we have AI tutors … It is a deeply human process that AI is hopelessly incapable of meeting in a meaningful way. It's like being a therapist or a nurse."[13] He argues that AI tools customized to augment teaching and teachers are likely to be more effective for such purposes than in being "personal tutors."[14]

In that regard, the Trentino Group cites the recent example of Khanmigo to evaluate the application of AI technologies in the classroom. Salman Khan was a hedge fund analyst who started giving distance learning lessons in maths to his young cousin Nadia. He made practice videos and slides which she soon shared with her friends; and he went from tutoring her to tutoring 15 kids, and many more. That became Khan Academy with an ever-greater offering of online courses and distance learning opportunities. In August 2020, Khan, his academy doing very well at the time, writes in the *New York Times* that distance learning was suboptimal but that, in the context of the pandemic, it was a necessity.[15] Talking about the value of the classroom experience and about the importance of interaction, he adds: "Because every child's life has become more distanced during the pandemic, there's an even higher burden on distance learning to emphasize human connection."[16] Cut to three years later, and Khan Academy has partnered with Open AI to roll out a tutorbot, called Khanmigo, an experiment that has been extensively reported on by the *New York Times*.

In this reporting, what we see is Khanmigo being used in maths classrooms in primary and middle schools in New Jersey in the US. *The Trentino Brief*'s authors note some interesting outcomes and problems from the Khan Academy case: First, that before a tutorbot becomes part of the classroom, as with any kind of automation, there is displacement onto the teacher to do other kinds of work to ensure that its answers are correct because LLMs make mistakes; also, the teacher must ensure that the technology is giving the student opportunities to actually work out problems by themselves rather than just give them the answers straight away. Teachers learn through experience how best to calibrate the time required for a group or individual to respond; sometimes, they need time to work it out, at other times, an answer that is wrong (or right) might be an opening for greater discussion. So, an unequal division opens between people who do the work of teaching in the classroom, because not only are they actually monitoring the system in its interactions with students, but they are also feeding back to the model in real time, improving it and, eventually, its value for the elites who own Open AI.

Second, a school pays US$60 per student for Khanmigo. They are not buying a product but are part of the rentier economy in which everything is assetized, i.e., it becomes an asset that gets rented out for access. In this model, tech companies like

13 Jeffrey R. Young, "A Technologist Spent Years Building an AI Chatbot Tutor: He Decided It Can't Be Done," *EdSurge*, January 22, 2024, https://www.edsurge.com/news/2024-01-22-a-technologist-spent-years-building-an-ai-chatbot-tutor-he-decided-it-can-t-be-done.

14 See Young, "A Technologist."

15 See Sal Khan, "I Started Khan Academy: We Can Still Avoid an Education Catastrophe," *The New York Times*, August 13, 2020, https://www.nytimes.com/2020/08/13/opinion/coronavirus-school-digital.html.

16 Khan, "I Started Khan Academy."

Open AI exact economic rent from educational institutions in the shape of ongoing subscriptions for digital services, and can derive further value from extracting data about usage too. The costs of running large language models are "eye-watering," to quote Sam Altman,[17] so it makes sense that tech companies would develop this sort of business model.

Third, *The Trentino Brief* cannot identify research about what this kind of educational model actually does to learning, to the cognitive functions of children and learners. There might be some great value to education bots, or generative AI in education, and there are also likely to be negative outcomes.

If learning through education bots is likely to be challenged, then the repository of humanities knowledge that is AI might remain just that for some time—a repository. Some members of the Trentino Group caution that there is a history to technology making silver-bullet-type promises to address social issues and problems; the infamous One Laptop Per Child (OLPC) program launched by Nicholas Negroponte, founder of MIT Media Lab, in 2005 is one that relates directly to education technologies. OLPC was an ambitious project to bring hand-cranked laptops (read: simple, analog, Global South-proof) to children in the Global South to "facilitate access to technology as a way to combat the educational gap."[18] "Infamous" because, as Morgan Ames finds in the case of Paraguay, the irresistible "charisma" associated with the notion of a transforming the lives of young boys in global South slums failed quite spectacularly; it simply did not work because the challenge of bringing education to the impoverished South emphasized the technology rather than the web of social relations that education takes place in.[19] Moreover, OLPC was expensive for local governments in the South to buy and host. David Souter writes about Negroponte's assumption that

> if you gave children laptops, they would teach themselves to do all kinds of things, leapfrog the adult world, become vectors of change for older generations and for whole societies. Teachers in this model were unnecessary, and OLPC did not provide a teacher interface or backup. The children got their laptops, were expected to learn with them, fix them when they went wrong, and change the world with them.[20]

OLPC assumed that transformative education and learning happen through "self-learning," and that schools were neither functional nor even necessary. Some members of the Trentino Group from the Global South grumble about testing in their countries and "dumping" of failed technologies from the North to the South. Group members from the North are more open to trying out things with generative AI in education in their countries, because, well, it would spur innovation and create new opportunities.

17 Sam Altman (@sama), "we will have to monetize it somehow at some point; the compute costs are eye-watering," Twitter, December 5, 2020, 8:38 a.m., https://twitter.com/sama/status/1599669571795185665.

18 "About OLPC," OLPC, accessed February 2, 2024, https://landing.laptop.org/aboutolpc/.

19 See Morgan G. Ames, *The Charisma Machine: The Life, Death, and Legacy of One Laptop per Child* (Cambridge, MA: MIT Press, 2019).

20 David Souter, "Inside the Digital Society: Lessons from Little Laptops," The London School of Economics and Political Science, January 13, 2021, https://blogs.lse.ac.uk/parenting4digitalfuture/2021/01/13/one-laptop-per-child/.

Better Data

In its quest to secure a future for the humanities *and* for AI, *The Trentino Brief* emphasizes, first and foremost, the quality of the data feeding AI needs to improve: Data scraped indiscriminately from the internet is of poor quality, and sometimes it contains offensive and illegal material.[21] Generative AI therefore continues to be discriminatory and biased,[22] and, furthermore, language models fabricate responses that may be incorrect. Another aspect to be considered is the fact that data labelling is low-paid work offshored to Kenya, the Philippines, and India as part of a well-established global data work industry.

The Trentino Brief thus proposes that there could be more creativity in sourcing high-quality data. For instance, tech companies could buy rights to the vast numbers of children's and young adult (YA) fiction and use this to train language models in style, tone, and sentiment. Authors who write for children and young people bring great degrees of care, creativity, and attention to their work, and the variety of dramatic styles applied in YA writing might offer generative AI more of a range to sample from. In the process, book authors, translators, and illustrators could see increased incomes.

But more is needed, say the *Trentino Brief*'s authors, than just improving data quality: what is required is its more precise reworking to improve tonality, rhetorical style, and delivery. People communicate differently on Reddit, on X, in marketing copy and in academic textbooks. This diversity of tone and voice is missing in LLMs; the default tone tends towards anodyne "marketing copy speak" to be inoffensive and palatable to a range of audiences. However, this flatness means that LLM users must actively rework the results to make it suitable for wider use. This is perhaps why Silicon Valley companies are already hiring poets and writers to bring a more "literary" quality to its outputs. Not only would it be of value for an LLM to generate distinct literary styles, but it would also introduce capabilities to build certain forms like poetry, song lyrics, and narrative writing.[23]

The *Trentino Brief*'s authors argue, rightly, that poets, literary scholars, translators, historians, cultural theorists, and performance artists understand language and its power deeply.—Who better to work at kneading and massaging generated language suitable for different contexts? Furthermore, there is a logic that mirroring forums like Mumsnet rather than Reddit might deliver better tonality and "voice." Mumsnet is a UK parenting advice website with millions of users. Aside from its content (which might range from the health benefits of goji berries, to keto diets, to keeping children safe from online grooming), what is interesting about it is the quality of its advice and how it is shared: clear, honest, heart-felt, direct, tried-and-tested, with very little snark, violence, or vulgarity. (Parents in the Trentino Group swear by Mumsnet.)

21 See David Thiel, "Investigation Finds AI Image Generation Models Trained on Child Abuse," Stanford Cyber Policy Center, December 20, 2023, https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse.

22 See Eryk Salvaggio, "Composites and Correlations: Media Representations in the Age of AI," *Cybernetic Forests*, February 26, 2023, https://cyberneticforests.substack.com/p/composites-and-correlations.

23 See Andrew Deck, "Why Silicon Valley's Biggest AI Developers Are Hiring Poets," *rest of world*, September 20, 2023, https://restofworld.org/2023/ai-developers-fiction-poetry-scale-ai-appen/#:~:text=Silicon%20Valley%20training%20data%20giants,quality%20of%20generative%20writing%20tools. See Deck, "Why Silicon Valley's Biggest AI Developers."

Combating Anthropomorphism

The *Trentino Brief*'s authors identify a role for humanities education and training in enhancing generative AI by attending to the language used. This is related to the issue of data quality but is quite specific: it is about the implications of anthropomorphic language for the emergence of AI. They propose a model of trained writers working in real time to attend to how LLMs communicate and how humans communicate about AI, particularly scientists, journalists, and writers. For their words will obviously go into language models and create impressions of AI, it is essential to have an oversight mechanism to ensure that society maintain appropriate ways of talking about AI.

Anthropomorphism is a common feature of how humans approach non-human others, such as in how we gender vehicles, give names to pets or digital assistant systems like Alexa, Jeeves, or Siri, and say "please" and "thank you" to language-generating agents. This runs the risk of us projecting onto machines capacities that do not exist as they do in us. (Here, the group cites the recent example of a Google engineer believing the LaMDA system to be "conscious" when what the term means in humans continues to elude and enchant scientists. Google fired the engineer and clarified that the system was not conscious.[24])

AI portrayed in the form of humanoid robots or automated superhumans has been actively advanced by scientists, science fiction authors, and tech companies.[25] It is believed that AI is "thinking" rather than "processing" or "performing computation." Significantly, technology is designed to elicit interaction and connection; it is a choice to give a technology a female name, to render a robot with humanoid features, or to have a large language model like ChatGPT use the phrase "How can *I* help you today?".

Humans anthropomorphize for a number of reasons: our need to establish social relations is foundational to being human, and so we establish trust and cooperation through voice and the eyes. Human-computer interaction (HCI) specialists and scientists who design collaborative robots investigate the relationship between human trust, and the presence of eyes, their location, and gaze, in non-human others.[26] The implications of anthropomorphism are that if we believe that AI systems are *like* humans or will *approach* humanlike qualities of mind, then we think AI is amenable to being treated *as if* human. We believe that "the AI" is responsible for error or harm rather than the human organizations and systems behind it.

Grappling with this legacy of anthropomorphism and its risks requires, so the Trentino Group concludes, that "rich psychological terms," such as awareness, perception or the idea that AI has a mind, are to be treated with caution.[27] Also, the group again stresses the fact that AI is constituted by infrastructures, people in organizations, laws, regulations, global code repositories, programmers, engineers,

24 See Nitasha Tiku, "The Google Engineer Who Thinks the Company's AI Has Come to Life," *The Washington Post*, June 11, 2022, https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/.

25 See "AI Narratives," Leverhulme Centre for the Future of Intelligence, accessed February 2, 2024, http://lcfi.ac.uk/projects/ai-narratives-and-justice/ai-narratives/.

26 See Artur Pilacinski et al., "The Robot Eyes Don't Have It. The Presence of Eyes on Collaborative Robots Yields Marginally Higher User Trust but Lower Performance," *Heliyon* 9, no. 8 (2023), https://doi.org/10.1016/j.heliyon.2023.e18164.

27 See Henry Shevlin and Marta Halina, "Apply Rich Psychological Terms in AI with Care," *Nature Machine Intelligence* 1, no. 4 (2019): 165–167, https://doi.org/10.1038/s42256-019-0039-y.

data harvesting mechanisms, mathematical reason, and sensors. As a consequence, people with education in the humanities could find data work in providing better language in how society talks about AI, managing anthropomorphism, biased language, discrimination, voice, tone, and so on.

RLHF

Many of the proposals the *Trentino Brief* outlines for improving AI and rescuing the humanities from its decline are tantamount to enhanced reinforcement learning with human feedback (RLHF), a remote data work practice for improving AI products. Instead of letting a generative AI system figure everything out on its own (which can lead to unexpected or unwanted behaviors), humans can provide feedback to it. This feedback could be in various forms: it might be direct instructions, corrections of the AI's actions, or even just humans ranking different AI-generated solutions or texts to show which ones are better. Thus, RLHF means that an AI is trained not just through trial and error on its own, but with the help of human feedback to guide it towards behavior and decisions that are desired.

One could think of CAPTCHAs that require human users to prove their humanity by identifying motorbikes, pedestrian crossings as a kind of RLHF; in this, humans "help" the computer vision in driverless cars to correctly identify objects in the real world, because this system does not "know" the difference between a motorbike and a lamppost and only "knows" the arrangement of pixels that constitute images of these objects.

RLHF is particularly useful when the tasks are complex, nuanced, or require a level of understanding that AI might not develop on its own. It is thus used in a variety of settings: from correcting the outputs of language models to robotics, from video game design and education technology to autonomous driving capabilities and healthcare. There are limits to RLHF, however, and, as the *Trentino Brief*'s authors state, companies working on generative AI must invest in mitigating the human biases that will affect this technique.[28] One way to enhance RLHF already integrated within the data work industry would be offering steady contractual employment to recent graduates in the humanities. Meaning, on a more positive note, that the remaining few humanities scholars and educators could pivot to becoming curators, critics, and facilitators in the new educational landscape, focusing on providing the human insight and experience that AI cannot replicate.

The Last Professor

While the authors of *The Trentino Brief* agree that AI has much to offer the humanities towards its next iteration and vice versa (given that creative practitioners and scholars with jobs in the academic humanities might use AI), the group also spent some time thinking about legacies and resistance. In that regard, the authors are divided on the maintenance of physical archives of artifacts associated with the humanities: It is possible that the experience of an everyday object or concept might be lost through successive layers of misinformation or translation through a machine system that does not really "know"

28 See Stephen Casper et al., "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback," preprint, last revised September 11, 2023, https://arxiv.org/abs/2307.15217.

anything; and equally, perhaps, through climate change or cultural erasure through societal change and conflict. While human knowledge gets digitized, should we maintain physical archives of them or not? Something like the Svalbard seed bank, perhaps?

Human cognition, language, experience, and thought are not universal but dependent on embodiment, place, and context. The humanities attempt to capture the complex realities of being human in a plural, often irrational, uncontrollable, and changing world. If future humans are to keep building this knowledge for themselves and for training future machines, then, who will maintain the remnants of a disappearing world and the taxonomies that will organize all this knowledge? Furthermore, if there needed to be fact-checking or an "original" source of verification of something—say, the taste of a banana or the authenticity of a piece of tapestry—then there might need to be a small community of people who will have to be supported to do this work. Thus, *The Trentino Brief* notes that a separate subcommittee for Future Human Archives might need to be set up to explore how this will be funded and maintained.

In addition, *The Trentino Brief* also attends to the possibilities of resistance from within the humanities to these efforts to fold their work into the shaping of future AI. The financial and social crises faced by humanities enthusiasts, academia, societies, and institutions will lead to crises of identity among these communities. How will they cope, readjust, or resist, and what will the cost of this be? What might its potential be? For instance, there might be an underground humanities movement in which underground "real humanities" clubs are organized for literature and art to be discussed and taught by actual humans, preserving the human element in these subjects.[29] There might be a surge in popularity of personal storytelling in entirely digital-free zones to minimize digitization as far as possible. Here, people might gather to share and listen to personal narratives, experiences, and interpretations of cultural artifacts. The Trentino Group acknowledges that there could be some positive outcomes but advise caution lest the plans for AI's improvement not achieve their targets.

Editor's note: The first installment of the *Brief*'s reconstruction ends here; further information about the final version of the text and the Trentino Group's discussions is hard to find at this stage. But the secretarial network might eventually piece together more details.

Afterword

In 1999, the French-American journalist Susan George published a parafictional work, *The Lugano Report: On Preserving Capitalism in the 21st Century*.[30] The report's fictional authors are a singular, multidisciplinary "Working Party," assembled to address the problem that shrinking resources, ecological devastation, and excessive consumption are negatively affecting the vibrance and potential of capitalism. The Working Party must craft a plan for how the winners of the current economic system can maintain their future. The report proposes that the only way forward in the interests of maintaining growth is that global populations must "reduce."

29 This was an idea generated by ChatGPT-4.

30 For further information on the book, see the article of the same name published by Susan George on July 5, 2005, on the Transnational Institute's homepage: https://www.tni.org/en/article/the-lugano-report.

Bluntly put: they propose what can only be slow genocide; there are just too many people on the planet, and most of them, the poor, will continue to be a burden on limited planetary resources. If those in power are to maintain their status, this is the only way through, concludes the Working Party.

Without its important afterword, *The Lugano Report* might be considered an actual proposal, one that approaches reality. When disasters happen, they tend to affect the poorest, registering global concern, often thanks to the rapid proliferation of news and images via social media and global news networks. However, concern does not necessarily transform into action. Disasters—and the marginal—are vulnerable to the actions of more powerful communities elsewhere. For instance, global teams of scientists working in the niche field of attribution science showed that the scorching heat wave that affected North India and Pakistan in 2022 was thirty times more likely to occur because of greenhouse gas emissions in the global North.[31] Similarly, the devastating floods that destroyed a third of Pakistan's farmland in 2022 was caused in large part by human, socioeconomic, and political factors such as "the proximity of human settlements, infrastructure (homes, buildings, bridges), and agricultural land to flood plains, inadequate infrastructure, limited ex-ante risk reduction capacity, an outdated river management system, underlying vulnerabilities driven by high poverty rates and socioeconomic factors (e.g. gender, age, income, and education), and ongoing political and economic instability."[32] The loss from the flooding in Pakistan came to US$15 billion, and the country requires US$16.3 billion in recovery aid.[33] According to Oxfam International, the world's billionaires continue to burn through the remaining carbon left on the planet, contributing to its rising temperature; if it rises beyond 1.5 degrees, the greatest negative impacts will be felt by women and girls, by the poor, and by indigenous people.[34] (Power likes to perpetuate itself.)

Susan George says she does not endorse any of the premises, outcomes, or methods employed by the fictional Working Party. And she *wants* people to be chilled and moved to fury by the text. She *wants* her readers to question the *premises* ("except for the ecological ones"[35]) of the *Report* rather than its conclusions or recommendations. "This book is intended to afflict the comfortable without, alas, providing much comfort to the afflicted," says George.[36] She creates a cast of characters of the Working Party: they are " 'policy intellectuals', the kind who switch effortlessly from academia to government and back, running prestigious university centers and acting as highly placed advisers."[37] She gives them pseudonyms and places them in a large, comfortable house in " 'neutral', i.e. Swiss, territory, at once charming, discreet and rich. Lugano sprang

31 See "Climate Change Made Devastating Early Heat in India and Pakistan 30 Times More Likely," World Weather Attribution, Imperial College London, May 23, 2022, https://www.worldweatherattribution.org/climate-change-likely-increased-extreme-monsoon-rainfall-flooding-highly-vulnerable-communities-in-pakistan/.

32 World Weather Attribution, "Climate Change."

33 See World Weather Attribution, "Climate Change."

34 See Asfaq Khalfan et al, *Climate Equality: A Planet for the 99%* (Oxford: Oxfam International, 2023), https://doi.org/10.21201/2023.000001.

35 Susan George, *The Lugano Report: On Preserving Capitalism in the 21st Century* (London: Pluto Press, 2003), Afterword, EPUB.

36 George, *The Lugano Report*, Afterword.

37 George, *The Lugano Report*, Afterword.

readily to mind."[38] As such, their intellectual portfolios have wide-ranging impact and influence on communities they are neither part of nor represent; in other words: the outcomes of their work will not affect them directly.

Satire is a vehicle for speech within political and social restrictions, such as under conditions of authoritarianism. At the same time, relatively open societies have their sacred cows too; satire can work as a thought experiment, presenting opportunities to articulate ideas that might be awkward, uncomfortable, or ahead of their time. George's *The Lugano Report* is a satirical work that acts as improvisational theatre does: enabling transformation, working out ideas, testing out new configurations. *The Trentino Brief* is concerned with the transformations taking place in higher education, epistemology, and pedagogy through the emergence of AI; and it is directly inspired by *The Lugano Report* in two ways. First, in how satire can offer opportunities for articulating controversial or unpopular ideas; and, second, in drawing attention to the social, economic, political and cultural worlds of industry, academia, and regulation where decisions about technology are made. As such, *The Trentino Brief* is both an essay (this one) and also a new container for curated and original articles, blog posts, and curricula in order to examine how AI is transforming formal institutions and practices of knowledge-making.

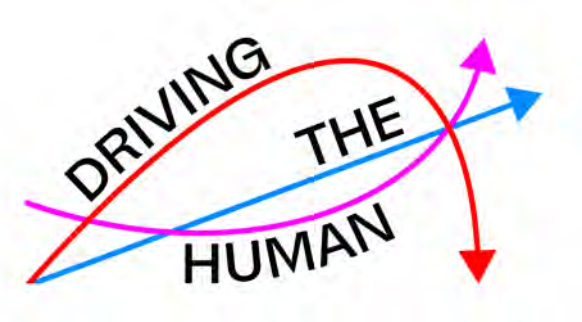
The Collingridge dilemma that animates *The Trentino Brief* is just one approach to the social life of technologies; there are many ways to think ourselves out of the horns of such dilemmas. For one thing, technologies develop incrementally, not in huge leaps. Jonnie Penn notes that four centuries passed between the invention of the technology and its eventual transformation of societies.[39] At the time of its invention, there was neither a supply of paper nor widespread literacy; these had to be developed through elaborate social transformation that took time, social, cultural, and political change.

By monitoring and adapting to small changes, regulators and developers can manage and mitigate risks more effectively, rather than waiting for complete understanding or widespread adoption. By analyzing how past innovations were integrated into society and what challenges they posed, policymakers and developers can better anticipate and address potential issues in new technologies. We can also change our approach to regulation and control, opting for more flexible and adaptable regulatory frameworks that can evolve alongside technological advancements: instead of static rules, agile regulation allows for continuous adjustment and efinement of policies as more is learned about a technology's impact.

Finally, the Collingridge dilemma can be managed by involving the public, by including non-experts in the discussion and decision-making process around new technologies like AI. How can everyday publics, who are subject to the development of technologies but have little say in the matter, think about adaptability in the face of change?—Rather than eliminating or avoiding uncertainty, more of us should be talking about the worlds we have and the worlds we want.

38 George, *The Lugano Report*, Afterword.

39 See https://www.cam.ac.uk/stories/cambridge-festival-2024-ai-technology.



Editorial team
Sarah Donderer (curator), Hannah Jung (curatorial assistant), Jens Lutz (project management)

Editing
Petra Kaiser

Graphic design
rapp.design, Leonie Rapp

Cooperation
Driving the Human has been initiated by Forecast, and further developed in continuous conversations between Freo Majer with Jan Boelen (Atelier LUMA), Martina Schraudner (Fraunhofer Center for Responsible Research and Innovation (CeRRI)), and the curatorial team of ZKM | Karlsruhe.

acatech: Sandra Fendl, Doerthe Winter-Berke, Hannah Lecheler

Karlsruhe University of Arts and Design/Bio Design Lab: Julia Ihls, Anthea Oestreicher, Hajo Eickbusch, Jehad Othman

ZKM | Center for Art and Media Karlsruhe: Peter Weibel (†), Sarah Donderer, Philipp Ziegler

Project coordinator: Nikola Joetze
Program coordinator: Vera Sacchetti
Production assistant: Sarah Lipszyc
Project assistant: Rabea Kaczor

A project by

FORE CAST

acatech
DEUTSCHE AKADEMIE DER TECHNIKWISSENSCHAFTEN

Staatliche Hochschule für Gestaltung Karlsruhe / Karlsruhe University of Arts and Design

ZKM Center for Art and Media Karlsruhe

Supported by

Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz

aufgrund eines Beschlusses des Deutschen Bundestages