Credit Where Credit Is Due: What Bhutan's intellectual and cultural patrimony offers to the global debate about Trustworthy AI

Stephan Sonnenberg[1]

From 2015 to 2020, I had the extraordinary privilege of living in Bhutan, having been tasked with helping launch the Jigme Singye Wangchuck School of Law, Bhutan's first and only law school.

During those five years, I spent a lot of time thinking about Bhutan's extraordinary intellectual footprint, including its development philosophy focused on the maximisation of Gross National Happiness (GNH) rather than Gross National Product, and how this concept influenced understandings of "justice" at the grassroots level.

In 2020, I moved to Korea, where I began teaching human rights at Seoul National University's School of Law. While there, I got involved in a project to document and engage with debates taking place at the United Nations about how best to think about new and emerging digital technologies (NEDTs), including artificial intelligence (AI), at the global level. Together with key contributing authors in Seoul and Geneva, and drawing on contributions from experts from around the world, we have proposed a practical, flexible, and human rights-based framework for governing these emerging digital technologies. (Seoul National University AI Policy Initiative, n.d.)

Throughout that work, I found myself returning frequently to lessons I learned during my time in Bhutan.

This short essay can be understood as a long-overdue footnote -- an effort to give credit where credit is due -- but it is also directed towards policymakers in Bhutan itself. As Bhutan grapples with the opportunities

1 Stephan Sonnenberg is a professor of law teaching at Wesleyan University in the United States. From 2015-2020 he taught with the Jigme Singye Wangchuck School of Law in Bhutan.

and risks associated with AI and other NEDTs, it should once again look to both outside and indigenous sources of wisdom to craft its own approach towards these new technologies.

Regarding those distinctly Bhutanese sources of wisdom, I would suggest two lessons in particular.

Before describing them, a few important caveats are necessary to avoid this description falling into the trap of over-generalisation and romanticism. First, just like any other place in the world, Bhutan is neither culturally homogenous nor is it a place without any policy or philosophical inconsistencies. The Bhutanese people, just like all other peoples I know of, disagree about politics, development, business ideals, religion, the importance of new technologies, and what makes for the "good life". Romanticising Bhutan as though it were some mythical "Shangri-la" antidote to the rest of the world's troubles would do a disservice to the very real diversity of thought and real-world policymaking innovations that have made Bhutan's voice so durably relevant in global policy discussions.

What follows are merely my personal takeaways from my personal journeys through Bhutan's intellectual landscape.

When thinking about the governance of AI, a focus on human rights is important, but human rights coupled with a sense of human responsibilities offer a far more comprehensive ethical framework

Human Rights have rightly emerged as perhaps the principal normative anchor for global discussions about the governance of AI and other NEDTs.

In its 2023 Interim Report, the United Nations High-Level Advisory Body on Artificial Intelligence concluded that "human rights must be at the centre of AI governance." (2024, p. 39) After extensive consultations with stakeholders around the world, that sentiment made its way directly into the text of the Global Digital Compact (United Nations General Assembly, 2024), which was adopted by the UN General Assembly in late 2024 to guide digital cooperation, internet governance, and the safe use of emerging technologies like artificial intelligence.

Human rights gives engineers and policymakers a common vocabulary to help identify those interests that should never be sacrificed in the name of technological progress: A commitment to equality and non-discrimination, privacy, the central importance of human dignity, freedom of expression, due process, political participation, and the right to an education, healthcare, and a clean environment, among several other such standard enumerated rights.

The introduction of AI and other potentially transformative technologies does not render that pre-existing framework obsolete. It does, however, highlight some of the blind spots that have always hampered the institutional, philosophical and legal edifice undergirding the human rights methodology, especially when it comes to addressing complex social human rights harms. (Kennedy, 2002)

Consider a hypothetical AI-based hiring system designed to support human resources officers combing through voluminous stacks of application materials. Imagine further that the underlying AI large language model (LLM) was trained on historical hiring patterns that themselves were skewed by longstanding patterns of societal discrimination (for example, societal racism, sexism, or caste, or geographically coded prejudices influencing past hiring decisions). (O'Neil, 2016)

Cognisant of this risk, technologists have developed numerous intervention strategies to correct for the data-bias problem.[2] Unfortunately, none of these

2 Those include (1) dataset selection and weighting (an extraordinarily laborious process of screening datasets to remove or down-weight certain problematic data, correcting mislabeled data, or introducing artificially created new and 'corrected' data); (2) choosing or eliminating certain data features a model can access to make decisions; (3) choosing between different AI-based model architectures (entailing tradeoffs between highly effective but less-interpretable systems, such as a complex neural networks, and more interpretable but less sophisticated AI models such as simple scoring systems); (4) fine tuning a larger model for a more specialised purpose, such as only summarising (but not ranking) job applications; (5) post-training reinforcement optimisation (involving human evaluators "reinforcing" certain types of model outputs based on pre-determined criteria for appropriateness); (6) application design (focusing not on the underlying model itself, but rather the use of that AI-output within an application making use of that AI-model, and in particular how that design influences subsequent human decision-making); (7) balancing the way modern retrieval systems function to generate relevant outputs in a way that does not devalue minority viewpoints or applications; (8) changing automated decision-thresholds and re-introducing human oversight for borderline cases to avoid taking unjustifiable action on algorithmically determined false-positives without careful human consideration; (9) instituting bright-line restrictions on deployments of any AI-based systems for certain high-risk applications (for example a system designed to automatically recommend sentencing terms for convicted criminals); and finally (10) designing effective periodic audits and redlining strategies to rigorously test whether models and applications that rely on those models have potentially discriminatory outputs.

strategies offers a surefire guarantee that the problem of bias embedded within a model's training data can be cured.

At the risk of inappropriately anthropomorphising AI systems, these various technological fixes are akin to advising a deeply biased person to simply refrain from saying racist things in certain situations because of the negative social repercussions, rather than encouraging them to fundamentally interrogate and ideally reconsider their racist worldviews.

These various corrective strategies do not actually eliminate the underlying biases; they only make AI models better at masking the underlying human biases that have been embedded deep within the layers of their model architecture. For a disturbing real-life example of the shortcomings of even one of the most sophisticated alignment strategies in the industry, one need look no further than the swarm breakout and deception strategies of display in the recent hack of Hugging Face by AI-agents trained on an OpenAI frontier model, the staggering details of which were still emerging as this article went to press. (Cotra, 2026)

Furthermore, some sources of bias receive a great deal more attention from the companies and technologists responsible for the models than others. Race and gender might be prominent examples of bias that receive a great deal of attention in the Global North, at least within AI ethics circles.

Other biases, however, might be less apparent to these same engineers and analysts, and consequently unaddressed by the various training and reinforcement learning processes described above. These biases could cause harm in places like Bhutan or other markets consisting of only a few hundred thousand consumers.

Returning to the example of the human resources application described above, how would such a system deal with instances of candidates with "unexplained" three-year gaps in their employment history, when they pursue meditation retreats, or instances where a candidate might insist on an "unusual" start date for a new job based on an auspicious astrological reading.

In any such scenario where an AI-based system causes disproportionate harm to a certain class of candidates (in this case, Bhutanese candidates),

the classic human rights vernacular would ask essentially three interrelated questions: (1) Which individuals' human rights have been violated, (2) by whom, and (3) what remedies might be available to those victims? (Roth, 2011)

This frame has been dubbed the Victim-Violator-Remedy model of human rights activism, and- even though many have critiqued it- it remains core to the vast majority of contemporary human rights campaigns. When it comes to AI, however, each of these three prongs begins to fray.

Regarding our example from above, who precisely would we specify to have been harmed by the above AI system? Surely some small number of candidates might have secured jobs but for the existence of this AI system, but many others might never have been competitive even without the biased AI-based screening system. Should they all nonetheless be eligible for the same remedy, and if so, how can the remedy be fit to the actual degree of harm suffered?

Finally, who within the "supply chain" of such an AI-system could be held responsible? Numerous actors may have been involved in bringing this hypothetical model to market: The company that originally developed the core LLM that the system on which the application was built (eg: OpenAI, Anthropic, or Mistral); the developer who built the application on top of that foundation model; the employer who purchased the AI-based application; the human resources department that decided how it would be used; or even the individual employee who actually used the model.

The recent settlement agreement against Meta in the United States illustrates how difficult it is to answer these questions using traditional rights-based logic. In order to settle a lawsuit brought against the company by various U.S. states alleging that it intentionally designed its social media applications to be addictive and potentially harmful to children, Meta agreed to make certain adjustments in the way its algorithm works for minors and pay up to USD $17.1 billion to fund youth mental health services and education in the U.S. (Kang & Tan, 2026).

But children in Bhutan (or Germany or Thailand or any other non-US territory, for that matter) will of course not benefit from those payouts even if they may also have suffered tangible harms. Moreover, some children in

the U.S. presumably suffered more significant harms as a result of using Meta's products than did others, and yet Meta's settlement makes no such targeted distinctions in terms of its remedies.

The rights-based system presumably does incentivise a company such as Meta to institute better safeguards to ensure its technologies do not violate the human rights of its users. But the system works clumsily and indirectly, leaving genuine room for improvement.

In Bhutan, the logic and language of human rights is supplemented and reinforced by an equally robust vernacular emphasising the importance of taking one's human responsibilities seriously.

By coupling a responsibilities-based logic with the underlying rights-based one, the operative question is no longer merely "What are victims entitled to expect from those who develop and deploy powerful new technologies?" but also "What should those who have the power to shape these new technologies proactively do for others' benefit, even when it would be impossible to bring a legal claim against them for not doing so?"

That second question matters.

Imagine an AI developer who discovers evidence that suggests, but does not conclusively prove, that a product she is responsible for structurally disadvantages a socially marginalised group. A legalistic approach might ask only whether the available evidence establishes a pattern of liability, whether a particular person can prove that she was injured, and whether applicable law requires corrective action. If any one of these questions can be answered in the negative there would be little need to take any corrective action.

A responsibility-oriented approach would begin instead with the developer asking herself whether she has the power and agency to perhaps ameliorate the situation in some tangible way based on what she has heard.

The difference is not between law and morality, or between "Western rights" and "Bhutanese responsibilities". Human rights law itself contains important positive duties to protect, prevent, and progressively realise so-called "positive" economic, social and cultural rights.

The distinction is instead one of ethical orientation. Rights ordinarily establish claims that people may assert against states or institutions and indirectly against one another. Responsibilities encourage us instead to consider our own ability to make the world a better place.

This insight resonates in a society like Bhutan's where ideas of interdependence and responsibility toward others -- including future generations and other sentient beings -- come easily to those steeped in local Bhutanese and Buddhist culture. But they are also endemic to other parts of the world. One need not necessarily believe in the concept of reincarnation, for example, to take seriously the idea that individual actions occur within a web of relationships and that our actions might have consequences far beyond what we individually may be able to predict.

For AI governance, this suggests several important shifts in emphasis.

The developer should ask not merely "Can I legally release this system?" but also "What responsibility do I assume by doing so?"

The company should ask not only "Can we monetise this technology?" but also "What obligations arise from possessing this capacity?"

And the individual user might ask not merely "Am I permitted to use this tool?" but also "Could there be unintended impacts of my use of this tool, even on people or communities I cannot even anticipate?"

Human rights remain core to such analyses, but the language of human responsibility prods us to go just a little bit further.

Human flourishing is more multidimensional than merely measuring aggregate economic growth

The second lesson I take with me from Bhutan has to do with Gross National Happiness (GNH). GNH is sometimes trivialised, especially by jaded non-Bhutanese analysts, as a sloppy attempt to measure whether people are literally "happy". In reality, the effort to pursue a GNH-based development strategy is far more sophisticated, and relevant well beyond Bhutan's borders.

For a long time, it has been clear to many scholars of development that societal progress cannot adequately be measured by counting economic activity alone. A society that grows its GDP but simultaneously becomes less socially cohesive, destroys its environment, suffers from growing wealth inequality, and loses its cultural vitality is a society that, by many definitions, has become more fragile and in some significant ways worse off than it was before.

Much contemporary analyses of AI focus on narrow and easy to quantify productivity metrics. These statistics have the capacity to move markets and influence geopolitics, but they are also based on an excessively narrow definition of societal "progress". Economic analysis in recent decades has also – sometimes grudgingly – come to embrace the importance of more subjective factors that influence human wellbeing, such as equity, human rights, environmental externalities, social capital, and the intangible value of culture.

But Bhutan is arguably a pioneering thought leader on such matters. It has learned valuable lessons on how best to institutionalise the idea that public policy should be evaluated against multiple dimensions of human flourishing, including but not limited to the aggregate wealth created by a given development initiative.

Suppose, for example, that an AI-enabled technology promises to dramatically improve agricultural productivity while also reducing the number of people required to cultivate a given acreage of land. A conventional analysis might ask only how much productivity gains might be expected, how much additional income might be generated, which workers might stand to lose their livelihoods, and whether those workers could be retrained to work in a different industry.

A GNH-inspired analysis would not abandon these questions, but it might also ask additional questions. For example, what happens to the vitality of the rural communities where this new AI-based model of agriculture will unfold? Would the introduction of this new technology encourage young people to remain in those communities, or would it inadvertently accelerate migration to Thimphu or abroad? Would the technology preserve or erode traditional sources of knowledge associated with vanishing occupations? Would it have an impact on families and relationships between neighbours?

What might be the environmental consequences of introducing such a technology? Would the new technology allow people to make better use of their time, or would it simply increase the expectations for further productivity?

In short, would the introduction of this new technology make the community more worth living in, or would it merely lead to greater profits for the individual entrepreneur who manages to introduce this technology first?

If these kinds of questions sound strikingly familiar to Bhutanese audiences, it is because they mirror precisely the GNH-based policy innovations that have defined policymaking in Bhutan for the past few decades.

These innovations require lawmakers to subject their narrowly tailored policies (say, a telecommunications bill) to a broader set of questions, not unlike those listed above, to ensure that the proposed policy meets not only the narrow objectives of the reform effort (i.e., improve telecommunications infrastructure in Bhutan), but also the broader set of considerations contained within GNH. At the macro level, the periodic GNH Surveys measured national progress towards the full achievement of GNH, and at the micro-level the GNH Policy Screening tool translated it into policy creation.

AI impact assessments are already becoming an important part of governance regimes globally. Bhutan could develop an approach to AI impact assessments and certifications that explicitly draws on GNH, requiring policymakers, developers, and deployers to consider the effects of AI and other NEDTs not just on economic productivity and human rights, but also on all nine domains of GNH.[3] Such an approach would not just symbolically invoke the values and virtues of GNH, but also operationalise it.

Credit Where Credit is Due

Neither of these ideas is exclusive to Bhutan. Analogues can be found in the philosophical, religious, economic and political traditions of many

3 Those are (1) psychological wellbeing, (2) health, (3) education, (4) time use, (5) cultural diversity and resilience, (6) good governance, (7) community vitality, (8) ecological diversity and resilience, and (9) living standards.

other societies around the globe, including some in the Global North. Nor does Bhutan have a ready-made answer to every question that may arise when trying to act on these principles.

What is true, however, is that Bhutan has spent an exemplary amount of political and intellectual energy exploring and acting on these ideals. The emerging global framework describing what makes AI or other NEDT's "trustworthy" asks whether AI systems are compatible with human rights priorities.

The conversation about Trustworthy AI would be significantly improved if it also asked, first, whether our chosen governance approach incentivises individuals and organisations to consider their unique responsibility to promote and protect the wellbeing of others, and second, whether we are using a sufficiently holistic approach as we evaluate the aggregate impact of an AI system.

Those two questions, I would suggest, are two significant contributions that Bhutan can bring to the global discourse on ethical AI governance.

References

Cotra, A. (2026, August 28). *The Hugging Face attack surprised me*. Planned Obsolescence. https://www.planned-obsolescence.org/p/the-hugging-face-attack-surprise

Kang, C., & Tan, E. (2026, August 26). Meta to pay up to $17.1 billion in landmark settlement over social media addiction claims. *The New York Times*.

Kennedy, D. (2002). The international human rights movement: Part of the problem? *Harvard Human Rights Journal, 15*, 101-125.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Roth, K. (2011, March 4). Defending women's human rights. Human Rights Watch. https://www.hrw.org/news/2011/03/04/defending-womens-human-rights

Seoul National University AI Policy Initiative (n.d.). *Towards a human rights-based approach to new and emerging digital technologies, including AI*. https://sapi.co.kr/towards-a-human-rights-based-approach-to-new-and-emerging-digital-technologies-including-ai/

United Nations General Assembly (2024). *Global Digital Compact (Annex to the Pact for the Future)*. United Nations. www.un.org

United Nations, High-Level Advisory Body on Artificial Intelligence (2024, September). *Governing AI for humanity* (p. 39). https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf