Gransche, Bruno*

Rethinking AI with Transformative Philosophy

Upskilling our technology views to guide the digital transition

https://doi.org/10.1515/cdbme-2024-2066

Abstract: This paper introduces the *Transformative Philosophy* (T:Phil) program, an initiative designed to embed ethical and socio-political literacy into the technological development process. T:Phil, developed within the *Digitalized Lifeworld* research cluster and in partnership with leading international technology reflection experts, provides a comprehensive framework for integrating philosophical insights with technical expertise. Aimed at decision-makers in the tech industry, T:Phil delivers tailored content through modular series such as *Re:Thinking Technology*, which challenge participants to reconsider their approaches to AI, digital infrastructures, and our language of technology.

The paper highlights two key modules of the T:Phil program: The first, *AI in an Atlas-View*, promotes a holistic understanding of AI by examining its technical, human, and environmental aspects. This module encourages participants to view AI as a complex socio-technical ensemble rather than merely a technical innovation. The second module, *AI as a Metaphor*, explores the metaphorical language used to describe AI, fostering critical awareness of how these metaphors shape our perceptions and influence the ethical considerations of AI deployment.

Through innovative delivery methods, including hybrid keynote-discussion events, short series, and intensive *Thinkathons*, T:Phil engages technical professionals in reflective practices that enhance their understanding of technology's – especially AI's – broader implications. The program aims to help cultivate a new cadre of ethically and philosophically informed tech experts capable of designing AI systems that promote equitable and just outcomes.

This paper briefly introduces the T:Phil program and provides examples of its application, demonstrating its potential to transform the way we integrate ethical reflection into technology development.

Keywords: transformative philosophy, ethics, AI, technology reflection, top executive development, training, ethical and socio-political literacy, technology metaphorization

1 Introduction

Understanding technology and our perceptions of it are vital when tackling the imminent ethical query: how to lead a fulfilling life in a world permeated by technology in every aspect. Our objective is to attain this goal by employing technology, while guaranteeing well-being as individuals in stable social relations, just institutions, and a nurturing environment. Crucial questions arise from this macro-perspective: What constitutes a good life? What conditions, skills, knowledge, and practices are indispensable? How can we ascertain preferable social relations, institutions, and environmental conditions? In this context, a significant inquiry arises: In what ways can technology, if at all, bolster a good life? How can we harness its potential without suffering its risks?

These intricate questions demand extensive multi-, inter-, and transdisciplinary research, along with expertise to arrive at adequately intricate solutions. AI is poised to revolutionize among others medicine and the broader health sector, thereby improving our lives in terms of well-being, longevity, and healthspan. However, to address the critical questions surrounding the delicate equilibrium between AI's beneficial potential and its inevitable risks or undesirable impacts, it is crucial to deepen our understanding of AI's essence (and its limitations), as well as the ways in which we perceive and conceptualize it.

***PD Dr. Bruno Gransche:** Karlsruhe Institute of Technology KIT, Institute of Technology Futures ITZ, Douglasstr. 24, 76131 Karlsruhe, Germany, e-mail: bruno.gransche@kit.edu

2 Putting gains in technology reflection to work

Philosophy, notably the philosophy of technology, along with disciplines such as the sociology of technology and Science and Technology Studies (STS), have a rich tradition of examining the complex interplay between technology, individual behaviour, societal structures, and power dynamics. Despite the wealth of insights generated by these fields, their existence and contributions remain largely unacknowledged by those spear-heading technological development and implementation. The rapid digitalization and proliferation of AI, however, are exposing the urgent need for non-technical frameworks, theoretical foundations, and ethical considerations to effectively guide the ongoing technological transformation.

To illustrate this point, consider the following question: Does a specific artifact (or technological object) possess an intrinsic or fixed value relationship? In other words, can artifacts or, by extension, technologies be categorized as inherently good or bad? Do technologies a) have morally relevant effects, b) maintain value neutrality and remain isolated from morality, or c) possess their own genuine morals (or moral sense), potentially qualifying them as artificial moral agents? The answers to these questions have significant implications for determining appropriate responsibilities and responses. These inquiries have been subjects of long-standing debate within the philosophy of technology, with some positions having been disproved or reconsidered considering emerging technologies such as highly automated systems and AI. While the notion of technology's moral neutrality (option b) has been widely rejected, the possibility of AI as an artificial moral agent (option c) is being revisited in the context of AI's rapid evolution and potential autonomy. This ongoing discourse underscores the importance of integrating the insights and expertise of technology-focused humanities disciplines in shaping our understanding and governance of AI.

By what kind of AI design are individuals' preferences, decisions, or actions facilitated or hindered? What intentions and side effects are associated with these designs? How do they interact with prevailing socio-political circumstances?

This case of the value-technology relationship serves as just one illustrative instance, underscoring the significant advancements in technology reflection and the urgent need to consider these insights (with continual updating, of course) in today's technology de-sign, development, usage, regulation, and beyond.

3 Bridging Engineering and Philosophy with Transformative Philosophy

In the context of integrated research, a significant challenge lies in the fact that technical disciplines at the vanguard of digitalization and AI-driven transformations often overlook the valuable discourses of the humanities (just as the technical state-of-the-art is frequently disregarded by humanities scholars). However, there is a growing need for ethical, socio-political, cultural, and epistemological literacy (which we could refer to as "technology reflection literacy") among technical experts, as well as a basic technical or digital/data literacy among humanities scholars and all citizens navigating a responsible life in an increasingly digitalized world.

While we recognize the importance of scientific and expertise specialization, it is not feasible for everyone to study everything. Therefore, it becomes crucial to identify and select relevant parts of reflective expertise that can and should be assimilated by technical experts. Due to time constraints, we cannot solely rely on reforming technical curricula (e.g., incorporating STS or ethics modules in data science master's programs) to educate future generations. Instead, we must forge connections between technology reflection and technology design in a manner that enables current professionals (not just students and future professionals) to incorporate highly relevant and widely accepted elements of technology reflection expertise as they shape our digitalized world. In essence, we require up-to-date, technically informed philosophers and up-to-date, philosophically/ethically informed tech experts to collectively design systems and AI that yield preferable options and lifeforms for the majority, if not all, while avoiding unacceptable outcomes for anyone.

3.1 The T:Phil Program

The "Transformative Philosophy" (T:Phil) program [1] seeks to address this need by facilitating inter- and transdisciplinary integrated research and development through cultivating these bridgeheads for reflective technology design. Developed within the cluster *Integrated Research* (funded by the BMBF) under the *Digitalized Lifeworld* cluster-part, the T:Phil program was conceived in collaboration with internationally recognized experts in technology reflection, including past, present, and future presidents of the Society for Philosophy and Technology SPT. The *Lifeforms in a Digitalized Lifeworld* research project, led by Bruno Gransche and conducted in part-

nership with the global infrastructure company VINCI Energies and the industry association VDE, laid the groundwork for the T:Phil program.

The T:Phil program's primary objective is to disseminate essential insights from technology reflection (especially philosophy and ethics of technology) – spanning from ancient Greek and Roman wisdom to cutting-edge international knowledge – to key actors in technology development fields. The main target audience includes decision-makers in tech companies and associations, ranging from mid-management to top executive levels, such as portfolio, digitalization, and innovation managers, CEOs (Chief Executive Officers), and CDOs (Chief Digital Officers).

3.2 The T:Phil Content

T:Phil offers a curated collection of content that presents the state-of-the-art in technology reflection on various topics, such as AI, digitalized lifeforms, infrastructures, competencies (up-, de-, and reskilling), digital literacy, metaphoric technology language, technology imaginaries, technology and time (durability, sustainability, repeatability, etc.), transparency, accountability, responsibility, explainability, trust, and more. The content is continually expanded and features a modular structure that can be tailored according to the genuine problem definitions of the target group itself. Instead of presuming to dictate the problems of technical experts, T:Phil reacts to the target group's problematizations by offering a customized set of tools, theories, and concepts. These modules are organized into interlinked series, each focusing on specific themes like ethics, technology imaginaries, language, and conceptual issues. One such series tested in the research project is the *Re:Thinking Series*, which encompasses various episodes such as *Re:Thinking Technology* (an introductory module of sorts), *Re:Thinking InfraStructure*, or *Re:Thinking TechMetaphor*. Each episode combines some of the aforementioned modules that can be accessed and explored independently of the series as well. Besides reacting to problem definitions, the T:Phil program also aims to change the habituated problem-perspective (e.g., recognizing bridges as potentially racist) by providing a curated set of 'provocations' or re-problematizations from a reflective point of view.

3.3 The T:Phil Methods

To effectively deliver the selected content (modules, series, and curated sets), the team developed a range of tailored formats (and is continuously refined) in collaboration with didactic experts, researchers from pedagogy and adult education, and specialists in personal development. While it would be ideal to engage in extensive "Philosophy of Technology" or "AI Ethics" courses spanning 30 hours per semester, such options already exist and are not widely utilized by the target group due to various time constraints and the need for more concise focus (which can also limit the scope of reflection practice). T:Phil, therefore, provides different tailored and tested options for the target group to engage with this program, accommodating various levels of time commitment, attention capacities, and media preferences. Examples include:

1. Hybrid online-onlife keynote-discussion events requiring no prior preparation.
2. A short series of three events combining
    a. sensitization/provocation, new perspectives, and alternative thinking on technology;
    b. applying irritation to the participants' daily work and discussing their observations in a group setting; and
    c. a half-day in-person workshop to discuss these observations using a tailored set of T:Phil tools and concepts in response to b).
3. For in-depth discussions, guided reflection, and practice/habituation of transformed thinking on selected subjects, a two-and-a-half-day closed-session in-person event called *Thinkathon* is available, allowing for more complex reflection (as demanded by the complexity of certain subjects).

Two examples of modules of the T:Phil program are as follows:

3.4 Example 1 "AI in an Atlas-View"

AI must be considered a complex socio-technical ensemble, not just as a technology or in purely technical terms. The term "AI" is imprecisely defined and can refer to various phenomena or only selected aspects of the overall phenomenon. AI has been characterized as a *science* (e.g., by Marvin Minsky of MIT), a *realization of behavior* (e.g., by Wolfgang Wahlster of DFKI), or *IT systems* (e.g., in the EU AI Act), among other descriptions. To responsibly engage with AI (in design, use, and regulation), we must view the entire landscape of involved aspects, akin to taking a bird's-eye view or satellite view, like an Atlas. Two such perspectives to consider are the a) *KI.Me.Ge. Atlas of AI* [2] and b) *Atlas of AI* from Kate Crawford [3]

a) The *KI.Me.Ge. Atlas of AI* provides an in-depth analysis and reflection of AI as a complex socio-technical ensemble,

focusing on legal accountability aspects, power structures, inequality, expectational, narrative, and emotional aspects, as well as AI's role as a specific technology imaginary and metaphor (see example 2 below).

b) The *Atlas of AI*, famously created by Kate Crawford, offers an extensive map-like summary called "Anatomy of an AI," depicting both technical and non-technical elements required for a single AI performance, such as Amazon's Alexa in Echo systems. Crawford's tagline, "AI is neither artificial nor intelligent," needs further clarification: 'AI is neither *merely* artificial nor *inherently* intelligent.' The anatomy map highlights AI's natural aspects, such as necessary IT raw materials, rare earths, energy consumption in server parks and data centers, and landfills where old IT parts are discarded. It also underscores the human roles involved, such as mining workers in Congo and China, e-waste landfill workers in India, Clickworkers in Pakistan, and everyday users who train AI systems by completing RECAPTCHAs. This Atlas view unveils AI's natural, human, cultural, political, and other non-technical aspects, enabling a comprehensive understanding of AI beyond its artificial and intelligent dimensions. It becomes clear that referring to AI only in terms of *artificial* and *intelligence* can be seen as an expression of a specific technology image that neglects the 'dirty' side of AI, such as landfills and mines. Responsible AI integration into society and everyday life must consider the entire picture, not just its polished marketing aspects.

3.5 Example 2 "AI as a Metaphor"

The term "artificial intelligence" represents a specific perspective, or a particular technology imaginary, as previously mentioned. Examining AI as a condensed metaphorization allows us to explore this concept further. A more explicit form of this metaphorization could be: 'Those IT systems are intelligent, like humans or other biologically intelligent entities.' In this metaphorical construction, the meaning of one concept (the source) is transferred to another (the target) without explicitly describing the operation. Metaphors enable us to understand a target concept in terms of the meaning aspects of a source concept.

In the case of AI, human intelligence (the source) involves encountering resistance to will and to freedom of action (as problems) and devising solutions to overcome this resistance and still achieve goals. Artificial intelligence (the target) signifies that IT systems can 'behave' or take steps to overcome resistance to their implemented goals, similar to humans. This perspective implies that understanding AI requires grasping the source concept (human intelligence) and identifying how it can be applied to IT, while recognizing additional properties (e.g., untiring) and excluding specific human attributes (e.g., emotions).

Using this metaphorization approach already enhances AI comprehension by avoiding the inappropriate transfer of source aspects to the target (like emotions, robot rights, or citizenship). Contemporary metaphor research (or metaphorology) further supports this understanding by revealing that even source meanings originate from past metaphorizations, necessitating an analysis to prevent the inadvertent transfer of inadequate meaning aspects. For instance, *intelligence*, initially referring to a human cognitive capacity, was metaphorically derived from a manual sorting capacity (from Latin *inter-leggere*), which involved sorting various items, such as stones and lentils. Similarly, autonomy, which refers to an individual's freedom to establish and adhere to personal maxims or norms, is now applied to artificial systems like cars and robots. However, autonomy was initially used for political freedom when describing the ability of nations to create norms and laws without external interference. Developing a metaphor awareness, particularly within technical discourses, allows individuals to avoid transference biases. By recognizing and reflecting on these metaphors, we can prevent the unintentional transfer of unsuitable meaning components from the source to the target (e.g., attributing consciousness to 'intelligent' IT systems simply because intelligent humans are conscious).

Author Statement

Research funding: The research Project *Lifeforms in a Digitalized Lifeworld* (LeDiLe) in which the first version of the Transformative Philosophy Program was developed and tested was funded by the Federal Ministry of Education and Research (BMBF) under the funding sign 16SV8677. Conflict of interest: The Author states no conflict of interest. Ethical approval: The conducted research is not related to either human or animal use.

References

[1] https://transformative-philosophy.com/, last accessed 2024-06-07.
[2] https://www.kimege.de/ki-atlas/, last accessed 2024-06-07.
[3] https://katecrawford.net/atlas, last accessed 2024-06-07.