

Why Artificial *Phrónēsis* might be Computationally Infeasible

Sebastian Buecker

Department of Philosophy, Technical University of Darmstadt
sebastian.buecker@tu-darmstadt.de

January 6, 2025

Abstract

AI systems, in all their variety, become more integrated into our *socially* structured life-forms, be it as automated driving cars, as caretaker robots or large language models (LLMs). As of today, such systems exhibit astonishing abilities to move around in our social lifeforms, albeit being far away from perfection or human like abilities. In order to reach human like abilities, especially with regard to moral judgements, the ideal of *artificial phrónēsis* has been suggested.¹ Aspiring towards this ideal would imply finding methods to implement the ability to judge in morally complex situations in AI systems, without having a specified rule to apply. Whether or not it is possible to reach this ideal and by which criteria we'd know this, remains an open question.² My Aim is to present a computationally infeasible problem that, in order to reach *artificial phrónēsis*, would need to be solved, arguing that if it were to be solved, the solution would need to constitute something *more* than just an algorithm.

My argument will proceed in two steps. (1) By building on Aristotle's conception in the *Nicomachean Ethics*, it will be shown that *eudaimonía*, as the end which orients a phronetic agent, can be viewed to be functionally analog to *consistency* and how considering *consistency* affects mathematical judgments. This analogy will concern only the *formal* aspects of how both influence judgments in their respective domain: both *eudaimonía* and *consistency* are no objectives attainable in time, both cannot be set as an external goal, but they orient their respective actions intensionally and lastly, both demand a "holistic" viewpoint. (2) By using this analogy it will be possible to transfer a computational problem from mathematical logic into the realm of phronetic judgements. Because of Gödel's incompleteness results,³ *consistency* cannot be proven from *within* the

1. C.f. John P. Sullins, "Artificial Phronesis: What It Is and What It Is Not," in *Science, Technology, and Virtues: Contemporary Perspectives*, ed. Emanuele Ratti and Thomas A. Stapleford (Oxford University Press, 2021) (hereafter cited as Sullins, *Artificial Phronesis*).

2. "AP raises the question of whether or not a functional equivalent to phronesis is something that can be programmed into machines." Sullins, *Artificial Phronesis*, p.138.

3. Kurt Gödel, "Über formal unentscheidbare Sätze der *Principia mathematica* und verwandter Systeme," in *Kurt Gödel Collected Works : Volume I. Publications 1929-1936*, ed. Solomon Feferman et al. (New York: Oxford University Press, 1986).

respective calculus, instead it needs *outside* considerations.⁴ This also constitutes a problem for any algorithmic realization of artificial *phrónēsis*, because whatever algorithms *can* do, needs to be formalizable *within* the bounds of such a calculus. Thus the “modus operandi” by which mathematicians judge a calculus to be consistent and also, using the analogy, by which a phronetic agent judges his actions to be in accord with *eudaimonía*, cannot be implemented as an algorithm.⁵

The fact that Gödel’s incompleteness results restrict the realizable abilities of all conceivable algorithms was already confirmed by Alan Turing in his seminal paper “Computing Machinery and Intelligence,”⁶ but at the same time, Turing remained sceptical whether we humans possess the ability to overcome the bounds discovered by Gödel. The argument sketched above only gives a negative reason which methods cannot constitute a phronetic agent. The question towards Aristotle thus is: how do we as humans develop the virtue of *phrónēsis*, when *all* aspects of a *téchnē* need to be ignored, and would it be possible to emulate this path to achieve *artificial phrónēsis*?

4. “Calculus” here means a set of axioms, from which theorems are deduced using first order logic. “Inside” then refers to possible deductions in this logical system, while “outside” is negatively determined as all arguments that are not reducible to “inside” deductions.

5. Thus, in accord with Aristotle, *phrónēsis* cannot be regarded as a *téchnē*.

6. Alan M. Turing, “Computing Machinery and Intelligence,” *Mind* 59 (236 1950): 433–460.