© THE NEW INSTITUTE / Maximilian Glas

Kit Fine, Ph.D.

Silver Professor of Philosophy and Mathematics

New York University

Born in 1946 in Farnborough, United Kingdom
Studied Philosophy, Politics, and Economics at Balliol College, Oxford

ARBEITSVORHABEN

A Numerical Model for Parity and Imprecision

Building on the work of Ruth Chang, I develop a numerical model for parity and imprecision in value and in belief. The basic idea is to replace the notion of an approximate value or an approximate difference in credence with the notion of an approximate difference in values and an approximate ratio of credences. Thus we can no longer properly speak of the value (even the approximate value) of an item or the credence (even the approximate credence) of a proposition, but should speak instead of the approximate difference in values or the approximate ratio of credences. This simple move turns out to be an extremely powerful device that allows us to go far beyond what is representable by means of approximate values or approximate credences. We also jettison the idea of representing parity or imprecision by a set of precise values or a set of precise credences. Sometimes there exists no underlying set of precise values or precise credences; but, even when they do exist, they do not constitute a useful tool for computing how parities or imprecisions should be resolved. The framework should admit of a wide range of applications, and among the applications we wish to consider is one to decision-making in AI, wherein the notion of an approximate difference or ratio can provide a useful tool for the machine to communicate with the user on how hard cases are to be resolved.

Recommended Reading

Fine, Kit. Modality and Tense: Philosophical Papers. Oxford: Clarendon Press, 2005.
—. Semantic Relationism. Malden, MA: Blackwell, 2007.
—. Vagueness: A Global Approach. New York: Oxford University Press, 2020.

DIENSTAGSKOLLOQUIUM, 18.06.2024

Das Paritätsmodell: Die Einbindung des Menschen in der KI-Entwicklung

Das KI-Alignment, d. h. die Ausrichtung von KI-Outputs auf wichtige menschliche Werte, ist wohl das bedeutsamste ungelöste Problem in der Forschung und Entwicklung von KI. Wir schlagen einen Ansatz zur Entwicklung von KI vor, der auf einem neuartigen philosophischen Verständnis menschlicher Werte beruht. Das vorgeschlagene Modell ist eine Alternative zu herkömmlichen Formen der KI-Entwicklung – sei es in Form von maschinellem Lernen oder in Form von symbolischen Systemen –, die dazu beitragen kann, eine Wertausrichtung zu bewirken. Der numerische Rahmen, der dem Modell zugrunde liegt, verwendet Näherungsdifferenzen und Näherungsquotienten und ermöglicht damit, dass Werte als rechnerische Objekte behandelt werden können, und zwar außerhalb des üblichen Rahmens der Entscheidungstheorie. Wir zeigen, wie das Modell eingesetzt werden kann, um die Kommunikation zwischen KI und ihren Nutzerinnen und Nutzern zu erleichtern, wenn es um schwierige Entscheidungen geht.

PUBLIKATIONEN AUS DER FELLOWBIBLIOTHEK

Fine, Kit (Cham,2023)

Kit Fine on Truthmakers, Relevance, and Non-classical Logic

https://kxp.k10plus.de/DB=9.663/PPNSET?PPN=1877510092

Outstanding Contributions to Logic ; 26

https://kxp.k10plus.de/DB=9.663/PPNSET?PPN=1877510092

Fine, Kit (Dordrecht [u.a.],2023)

A semantics for the impure logic of ground

https://kxp.k10plus.de/DB=9.663/PPNSET?PPN=1853598283

Fine, Kit (London,2020)

The identity of social groups

https://kxp.k10plus.de/DB=9.663/PPNSET?PPN=1853595195