MIT Press • Open Encyclopedia of Cognitive Science

The Free Energy Principle

Maxwell J. D. Ramstead[1]

[1]**Wellcome Centre for Human Neuroimaging, University College London, London, UK**

MIT Press

Published on: Jul 24, 2024

DOI: https://doi.org/10.21428/e2759450.e4bf505d

License: Creative Commons Attribution 4.0 International License (CC-BY 4.0)

The free energy principle is a mathematical principle that describes how interacting objects or "things" (defined in a specific way) change or evolve over time. In this context, a *thing* is a set of states that can be meaningfully distinguished from other such things (e.g., particle, person, or population), both in the sense that there exists a boundary that separates that thing from other things, and also, through which it is able to interact with other such things. The free energy principle states that things, so defined (as sets of states that are separable from—but coupled to—other things) will look as if they track each other, by virtue of being separable but coupled. That things "track" each other means that they look to an external observer as if they carry information (or encode probabilistic beliefs) about each other. Mathematically, the free energy principle allows researchers to write down equations that describe the change of systems over time. This leads to a new family of mechanics, akin to classical and quantum mechanics, called Bayesian mechanics. In Bayesian mechanics, the time evolution of things minimizes a quantity called variational free energy, which is a function of probabilistic beliefs about the world that can be read in terms of self-information, *surprisal*, in information theory or Bayesian model evidence, *marginal likelihood*, in statistics. In summary, the free energy principle says that if something exists, then its dynamics must minimize variational free energy.

History

The free energy principle was originally proposed in the mid-2000s, in the context of theoretical neurobiology and computational neuroscience (Friston, 2005). It was first used to model the function, structure, and dynamics of the human brain. In this context, the free energy principle says that the brain is or entails a statistical model of its world and that brain and world synchronize via their interactions: In perception and learning, part of the agent (i.e., its brain and body) becomes more like the world; and in action, the world becomes more like the agent (Friston, 2010). Thus, the free energy principle allows researchers to describe how the brain responds to environmental perturbations. In this context, the free energy principle forms the basis of an approach to modeling intelligent systems—in particular, to modeling neuronal and living systems, such as brains and cell ensembles—called active inference (Parr et al., 2022).

Since being proposed in theoretical neurobiology, the domain of application of the free energy principle has been extended from neuroscience (Friston, 2010; Friston et al., 2021; Isomura et al., 2023). From the 2010s onward, the free energy principle was developed more generally as a mathematical principle of information physics (Friston et al., 2023a). It has now been applied to phenomena in domains as diverse as immune system function (Bhat et al., 2021), morphogenesis and pattern formation (Friston et al., 2015; Kuchling et al., 2020), evolutionary and developmental dynamics (Friston et al., 2023b), and the spread of information on social networks (Albarracin et al., 2020). More recently, the free energy principle has been deployed as a design principle for robotics and artificial intelligence (Çatal et al., 2021; Friston et al., 2024).

Core concepts

The free energy principle says something fundamental about what it means to be a physical thing that exists—in the sense that, through and despite its interactions with other physical things, that thing remains over time the kind of thing that it is: Quantum states, rocks, trees, and societies are things in this specific sense (Ramstead et al., 2023). The free energy principle says that any thing that exists (in this well-defined sense) will look to an external observer as if it minimizes a quantity called surprisal, or a mathematical upper bound on surprisal, called variational free energy [15]. Surprisal, simply put, expresses the improbability of some outcome, state, path, or configuration: It is the negative log of a probability. The free energy principle thus says that things exhibit behavior that is unsurprising, given the kind of thing that they are.

The free energy principle says that the motion of things in the space of possible states (or paths) minimize surprisal—in the same way that classical systems follow paths of least action that minimize an energy function (called a Lagrangian), which tells researchers how probable it is to find the system in a given state or configuration. This is a mathematical way of naturalizing the idea that things behave in unsurprising or characteristic ways, given the kind of thing that they are. The characteristic states or paths are described in terms of a generative model, which is a joint probability density over the states (or paths) that comprise the thing in question and states external to the thing. This rests on a stipulative definition of a thing in terms of its internal states and their boundary (known technically as a *Markov blanket*), where boundary or blanket (i.e., active and sensory) states mediate the coupling between internal and external states [16] . Thus, the free energy principle provides researchers with a stipulative definition of thingness, in terms of dynamical or causal coupling among states, and serves as the foundation of a calculus of belief updating that describes this coupling.

Applications of the free energy principle are twofold: First, it allows for the identification of the generative model—and implicit belief updating—that best explains the observable behavior of some thing. Second, it allows for the creation or stimulation of agents by solving for the paths of least action, under any given generative model of characteristic states (or paths).

Questions, controversies, and new developments

The epistemological status of the free energy principle—and in particular, the fact that it is not falsifiable—has been the subject of some controversy (Ramstead et al., 2023). The free energy principle is a piece of mathematical reasoning: It is no more subject to empirical falsification than other areas of mathematics, such as calculus. Instead, the free energy principle is used to write down specific generative models (and attending equations of motion), which are, in turn, subject to empirical verification (as in Isomura et al., 2023; see also Andrews, 2021). That real systems conform to variational principles like the free energy principle lends credibility to the free energy principle as applicable to specific classes of systems.

Broader connections

The free energy principle is closely related to predictive coding (Rao & Ballard, 1999) and the Bayesian brain (Knill & Pouget, 2004) in neuroscience [see Bayesian Models of Cognition]. The free energy principle has close ties to variational approaches in machine learning based on maximizing an evidence lower bound, which pick out exactly the same quantity as variational free energy (Friston, 2010). Finally, the free energy principle is also closely related to the principle of maximum entropy: The free energy principle turns out to be a way of writing down the principle of maximum entropy, under the constraint that systems maximize the accuracy of their predictions (Ramstead et al., 2023; Sakthivadivel, 2022).

Acknowledgments

The author is grateful to Karl Friston, Mahault Albarracin, and Axel Constant for helpful feedback.

Further reading

- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787
- Friston, K., Da Costa, L., Sajid, N., Heins, C., Ueltzhöffer, K., Pavliotis, G. A., & Parr, T. (2023). The free energy principle made simpler but not too simple. *Physics Reports*, *1024*, 1–29. https://doi.org/10.1016/j.physrep.2023.07.001
- Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: The free energy principle in mind, brain, and behavior*. MIT Press. https://doi.org/10.7551/mitpress/12441.001.0001
- Ramstead, M. J., Sakthivadivel, D. A. R., Heins, C., Koudahl, M., Millidge, B., Da Costa, L., Klein, B., & Friston, K. J. (2023). On Bayesian mechanics: A physics of and by beliefs. *Interface Focus*, *13*(3), Article 20220029. https://doi.org/10.1098/rsfs.2022.0029

References

- Albarracin, M., Demekas, D., Ramstead, M. J., & Heins, C. (2022). Epistemic communities under active inference. *Entropy*, *24*(4), Article 476. https://doi.org/10.3390/e24040476 ↩
- Andrews, M. (2021). The math is not the territory: navigating the free energy principle. *Biology & Philosophy*, *36*(3), Article 30. https://doi.org/10.1007/s10539-021-09807-0 ↩
- Bhat, A., Parr, T., Ramstead, M., & Friston, K. (2021). Immunoceptive inference: why are psychiatric disorders and immune responses intertwined? *Biology & Philosophy*, *36*(3), Article 27. https://doi.org/10.1007/s10539-021-09801-6

↩

- Çatal, O., Verbelen, T., Van de Maele, T., Dhoedt, B., & Safron, A. (2021). Robot navigation as hierarchical active inference. *Neural Networks*, 142, 192–204. https://doi.org/10.1016/j.neunet.2021.05.010

↩

- Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B, Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

↩

- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

↩

- Friston, K. J., Fagerholm, E. D., Zarghami, T. S., Parr, T., Hipólito, I., Magrou, L., & Razi, A. (2021). Parcels and particles: Markov blankets in the brain. *Network Neuroscience*, *5*(1), 211–251. https://doi.org/10.1162/netn_a_00175

↩

- Friston, K. J., Ramstead, M. J. D., Kiefer, A. B., Tschantz, A., Buckley, C. L., Albarracin, M., Pitliya, R. J., Heins, C., Klein, B., Millidge, B., Sakthivadivel, D. A. R., St Clere Smithe, T., Koudahl, M., Tremblay, S. E., Petersen, C., Fung, K., Fox, J. G., Swanson, S., Mapes, D., & René, G. (2024). Designing ecosystems of intelligence from first principles. *Collective Intelligence*, *3*(1), Article 26339137231222481. https://doi.org/10.1177/26339137231222481

↩

- Friston, K., Da Costa, L., Sajid, N., Heins, C., Ueltzhöffer, K., Pavliotis, G. A., & Parr, T. (2023a). The free energy principle made simpler but not too simple. *Physics Reports*, *1024*, 1–29. https://doi.org/10.1016/j.physrep.2023.07.001

↩

- Friston, K., Friedman, D. A., Constant, A., Knight, V. B., Fields, C., Parr, T., & Campbell, J. O. (2023b). A variational synthesis of evolutionary and developmental dynamics. *Entropy*, *25*(7), Article 964. https://doi.org/10.3390/e25070964

↩

- Friston, K., Levin, M., Sengupta, B., & Pezzulo, G. (2015). Knowing one's place: A free-energy approach to pattern regulation. *Journal of the Royal Society Interface*, *12*(105), Article 20141383. https://doi.org/10.1098/rsif.2014.1383

↩

- https://doi.org/10.1007/s10539-021-09787-1 ↩
- https://doi.org/10.3390/e14112100 ↩
- Isomura, T., Kotani, K., Jimbo, Y., & Friston, K. J. (2023). Experimental validation of the free-energy principle with in vitro neural networks. *Nature Communications*, *14*(1), Article 4547. https://doi.org/10.1038/s41467-023-40141-z ↩
- Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719. https://doi.org/10.1016/j.tins.2004.10.007 ↩
- Kuchling, F., Friston, K., Georgiev, G., & Levin, M. (2020). Morphogenesis as Bayesian inference: A variational approach to pattern formation and control in complex biological systems. *Physics of Life Reviews*, *33*, 88–108. https://doi.org/10.1016/j.plrev.2019.06.001 ↩
- Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: The free energy principle in mind, brain, and behavior*. MIT Press. https://doi.org/10.7551/mitpress/12441.001.0001 ↩
- Ramstead, M. J., Sakthivadivel, D. A. R., Heins, C., Koudahl, M., Millidge, B., Da Costa, L., Klein, B., & Friston, K. J. (2023). On Bayesian mechanics: A physics of and by beliefs. *Interface Focus*, *13*(3), Article 20220029. https://doi.org/10.1098/rsfs.2022.0029 ↩
- Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. https://doi.org/10.1038/4580 ↩
- Sakthivadivel, D. A. R. (2022). *Towards a geometry and analysis for Bayesian mechanics*. arXiv. https://doi.org/10.48550/arXiv.2204.11900 ↩