Emergent Causality & the Foundation of Consciousness

Michael Timothy Bennett[1][0000−0001−6895−8782]

The Australian National University michael.bennett@anu.edu.au
http://www.michaeltimothybennett.com/

Abstract. To make accurate inferences in an interactive setting, an agent must not confuse passive observation of events with having participated in causing those events. The "do" operator formalises interventions so that we may reason about their effect. Yet there exist at least two pareto optimal mathematical formalisms of general intelligence in an interactive setting which, presupposing no explicit representation of intervention, make maximally accurate inferences. We examine one such formalism. We show that in the absence of an operator, an intervention can still be represented by a variable. Furthermore, the need to explicitly represent interventions in advance arises only because we presuppose abstractions. The aforementioned formalism avoids this and so, initial conditions permitting, representations of relevant causal interventions will emerge through induction. These emergent abstractions function as representations of one's self and of any other object, inasmuch as the interventions of those objects impact the satisfaction of goals. We argue (with reference to theory of mind) that this explains how one might reason about one's own identity and intent, those of others, of one's own as perceived by others and so on. In a narrow sense this describes what it is to be aware, and is a mechanistic explanation of aspects of consciousness.

Keywords: causality · theory of mind · general intelligence.

1 Introduction

An agent that interacts in the world cannot make accurate inferences unless it distinguishes the passive observation of an event from it having intervened to cause that event [1, 2]. For example, say we had two variables, $R \in \{true, false\}$ and $C \in \{true, false\}$, where:

$$C = true \leftrightarrow \textit{"Larry put on a raincoat"}$$
$$R = true \leftrightarrow \textit{"It rained"}$$

Assume we have seen it rain only when Larry had his raincoat on, and he has only been seen in his raincoat during periods of rain. Based on these observations, the conditional probability of it raining if Larry is wearing his raincoat is $p(R = true \mid C = true) = 1$. A naive interpretation of $p(R = true \mid C = true) = 1$ is that we can make it rain by forcing Larry to wear a raincoat, which is absurd. When we intervene to make Larry wear a raincoat, the

event that takes place is not *"Larry put on a raincoat"* but actually *"Larry put on a raincoat because we forced him to"*. It is not that Bayesian probability is wrong, but interactivity complicates matters. The "do" operator [3, 4] resolves this so that $do[C = true]$ represents the intervention *"Larry put on a raincoat because we forced him to"*. It allows us to express notions such as $p(R = true \mid do[C = true]) = p(R = true) \neq p(R = true \mid C = true) = 1$, which is to say that intervening to force Larry to wear a raincoat has no effect on the probability of rain, but passively observing Larry put on a raincoat still indicates rain with probability 1. To paraphrase Judea Pearl, one variable causes another if the latter listens for the former [1]. The variable R does not listen to the C. C however does listen to R, meaning to identify cause and effect imposes a hierarchy on one's representation of the world (usually represented with a directed acyclic graph). This suggests that, if accurate inductive inference is desired, we must presuppose something akin to the *do* operator. Yet there exist at least two pareto optimal mathematical formalisms of general intelligence in an interactive setting which, given no explicit representation of intervention, make maximally accurate inferences [5, 6, 7, 8] . Given that the distinction between observation and intervention is necessary to make accurate inductive inferences in an interactive setting, this might seem to present us with a contradiction. One cannot accurately infer an equivalent of the *do* operator if such a thing is a necessary precondition of accurate inductive inference. We resolve this first by showing that we can substitute an explicit *do* operator with variables representing interventions. Then, using one of the aforementioned formalisms, we argue that need to explicitly represent intervention as a variable only arises if we presuppose abstractions [9]. If induction does not depend upon abstractions as given [6], then abstractions equivalent to the *do* operator may emerge through inductive inference. Beyond distinguishing passive observation from the consequences of one's own interventions, these emergent abstractions can also distinguish between the interventions and observations of others. This necessitates the construction of abstract identities and intents. We suggest this a mechanistic explanation of awareness, in a narrow sense of the term. By narrow we mean aspects of functional, access, and perhaps even phenomenal consciousness, and only if the latter is defined as "first person functional consciousness" [10] in the sense discussed by Boltuc in The Engineering Thesis on Machine Consciousness [11]; recognising phenomenal content such as light, sound, movement and so forth with one's body at the centre of it all [12]. To limit scope, we do not address the "hard problem of consiousness"[13].

2 Additional background

This section introduces relevant background material. The reader may wish to skip ahead to section 3 and refer here as needed. We present arguments rather than mathematical proofs, and the paper should be understandable without delving too deeply into the math. While all relevant definitions are given here, context is provided by the papers in which these definitions originated [6, 7].

To those more familiar with the agent environment paradigm, how exactly the aforementioned formalism represents cognition may seem unclear. Neither agent nor environment are defined. This is because it is a formalism of enactivism [14], which holds that cognition extends into and is enacted within the environment. What then constitutes the agent is unclear. In light of this, and in the absence of any need to define an agent absent an environment, why preserve the distinction? Subsequently, the agent and environment are merged to form a task [6], which may be understood as context specific manifestations of intent, bearing some resemblance to snapshots of "Being-in-the-world" as described by Heidegger [15]. In simpler terms, this reduces cognition to a set of decision problems concerning finite sets [16]. One infers a model from past interactions, and then makes a decision based upon that model (akin to a supervised learner fitting a function to labelled data, then using that to generate labels for unlabelled data). Arguments as to why only finite sets are relevant are given elsewhere [17, pp. 2].

2.1 Referenced definitions

Definition 1 (states of reality). *A set H, where:*

- *We assume a set Φ whose elements we call **states**, one of which we single out as the **present state** of reality.*
- *A **declarative program** is a function $f : \Phi \to \{true, false\}$, and we write P for the set of all programs. By **objective truth** about a state ϕ, we mean a declarative program f such that $f(\phi) = true$[1].*
- *Given a state $\phi \in \Phi$, the **objective totality** of ϕ is the set of all objective truths $h_\phi = \{f \in P : f(\phi) = true\}$.*
- *$H = \{h_\phi : \phi \in \Phi\}$*

Definition 2 (implementable language). *A triple $\mathcal{L} = \langle H, V, L\rangle$, where:*

- *H is reality, the set containing all **objective totalities**.*
- *$V \subset \bigcup_{h \in H} h$ is a finite set, named the **vocabulary**.*
- *$L = \{l \in 2^V : \exists h \in H \ (l \subseteq h)\}$, the elements of which are **statements**[2]*

(Truth) If we have a statement $l \in L$, and the totality of the present state of reality is $h \in H$, then l is **true** if $l \subset h$.

(Extensions) The **extension of a statement** $a \in L$ is $Z_a = \{b \in L : a \subseteq b\}$, while the **extension of a set of statements** $A \subseteq L$ is $Z_A = \bigcup_{a \in A} Z_a$.

(Notation) Lower case letters s, d, m, z, c represent statements, and upper case S, D, M, Z represent sets of statements. The capital letter Z with a subscript indicates the extension of whatever is in the subscript. For example the extension of a statement a is Z_a, and the extension of a set of statements A is Z_A.

Definition 3 (task). *Given language $\langle H, V, L\rangle$, a task is $T = \langle S, D, M\rangle$ where:*

[1] e.g. one may interpret declarative programs as activations in a neural network[7].
[2] Statements are formalised explicitly to avoid the symbol grounding problem [9].

- $S \subset L$ *is a set of statements called* ***situations****, where the extension* Z_S *of* S *is the set of all* ***possible decisions*** *which can be made in those situations.*
- $D \subset Z_S$ *is the set of* ***correct decisions*** *for this task.* [3]
- $M \subset L$ *is the set of all valid* ***models*** *for the task, where*

$$M = \{m \in L : Z_S \cap Z_m \equiv D, \forall z \in Z_m \ (z \subseteq \bigcup_{d \in D} d)\}$$

(How a task is completed) *Assume we have a hypothesis* $\boldsymbol{h} \in L$*:*

1. *we are then presented with a situation* $s \in S$*, and*
2. *we must select a decision* $z \in Z_s \cap Z_{\boldsymbol{h}}$*.*
3. *If* $z \in D$*, then the decision is correct and the task completed. This occurs if* $\boldsymbol{h} \in M$[4]*.*

Definition 4 (probability of a task). *Let* Γ *be the set of all tasks given an implementable language* $\mathcal{L}$*. There exists a uniform distribution over* Γ*.*

Definition 5 (generalisation). *Given two tasks* $\alpha = \langle S_\alpha, D_\alpha, M_\alpha \rangle$ *and* $\omega = \langle S_\omega, D_\omega, M_\omega \rangle$*, a model* $m \in M_\alpha$ *generalises to task* ω *if* $m \in M_\omega$*.*

Definition 6 (child-task and parent-task). *A task* $\alpha = \langle S_\alpha, D_\alpha, M_\alpha \rangle$ *is a child-task of* $\omega = \langle S_\omega, D_\omega, M_\omega \rangle$ *if* $S_\alpha \subset S_\omega$ *and* $D_\alpha \subseteq D_\omega$*. This is written as* $\alpha \sqsubset \omega$*. If* $\alpha \sqsubset \omega$ *then* ω *is then a parent of* α*, and* α *is a child of* ω*.*

Definition 7 (weakness). *The weakness of* $l \in L$ *is* $|Z_l|$*.*

Definition 8 (induction). $\alpha = \langle S_\alpha, D_\alpha, M_\alpha \rangle$ *and* $\omega = \langle S_\omega, D_\omega, M_\omega \rangle$ *are tasks such that* $\alpha \sqsubset \omega$*. Assume we are given only the definition of* α *and the knowledge that* $\alpha \sqsubset \omega$*. We are not given the definition of* ω*. The process of induction is:*

1. *Form a hypothesis by computing* $\mathbf{h} \in \arg\max_{m \in M_\alpha} |Z_m|$[5]*.*
2. *If* $\mathbf{h} \in M_\omega$*, then we have generalised from* α *to* ω*.*

2.2 Premises

There are two results pertaining to the formalism which we will adopt as premises:

> **(prem. 1)** "Weakness is [...] sufficient to maximise the probability that induction results in generalisation from α to ω." [7, prop. 1]

> **(prem. 2)** "To maximise the probability that induction results in generalisation from α to ω, it is necessary to weakness" [7, prop. 2]

[3] Note that each $d \in D$ is a superset of a member of S. S may be understood as a set of inputs, and D as the set of all unions of input and output which are correct.

[4] Note that $\forall m \in M : D \equiv Z_S \cap Z_m$, which means any $m \in M$ can be used to obtain D from S, because $D = \{z \in Z_m : \exists s \in S \ (s \subset z)\}$.

[5] Maximising weakness.

In other words, because this formalism employs weakness as a proxy, it maximises the accuracy of inductive inference (the probability that the model on infers will generalise[6]). For our purposes, this optimality less important than the representation of interventions it implies. In any case these results are not limited to lossless representations or optimal performance. Approximation may be achieved by selectively forgetting outliers[7], a parallel to how selective amnesia [18] can help humans reduce the world to simple dichotomies [19] or confirm preconceptions [20]. Likewise, a task expresses a threshold beyond which decisions are "good enough" [21]. The proof of optimality merely establishes the upper bound for generalisation. As a third premise, we shall paraphrase Judea Pearl, and require the emergence or presupposition of representations of intervention:

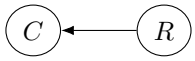
> **(prem. 3)** To make accurate inductive inferences in an interactive setting, an agent must not confuse the passive observation of an event with having intervened to cause that event. [1]

3 Emergent Causality

The formalism does not presuppose an operator representing intervention. Given our premises, we must conclude from this that either the formalism is not optimal, or induction as defined in definition 8 will differentiate between passive observation of an event and having intervened to cause that event.

3.1 The *do* operator is a variable in disguise

In the introduction we discussed an example involving binary variables R (rain) and C (coat). Typically the relationship between these two variables would be represented as a directed acyclic graph:

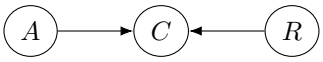


The intervention $do[C = c]$ deletes an edge (because rain can have no effect on the presence of a coat we've already forced Larry to wear) giving the following:

However we suggest this is a complicated way of thinking about the problem. The intervention can instead be represented by a variable A such that $p(C = true \mid A = true) = 1$ and $p(C = false \mid A = false) = 1$:

[6] Generalisation may be understood as correct labelling of out of domain data.

[7] For example, were we trying to generalise from α to ω (where $\alpha \sqsubset \omega$) and knew the definition of α contained misleading errors, we might selectively forget outlying decisions in α to create a child $\gamma = \langle S_\gamma, D_\gamma, M_\gamma \rangle$ (where $\gamma \sqsubset \alpha$) such that M_γ contained far weaker hypotheses than M_α.



This shows how accurate inference is possible in the absence of an operator representing intervention - just introduce a variable to represent each intervention.

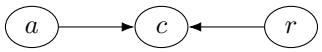
3.2 Emergent representation of interventions

This does not entirely resolve our problem. Even if intervention is represented as a variable, that variable must still be explicitly defined before accurate induction can take place. It is an abstract notion which is presupposed. Variables are undefined in the context of definitions 1, 2 and 3 for this very reason. Variables tend to be very abstract (for example, "number of chickens" may presuppose both a concept of chicken and a decimal numeral system), and the purpose (according to [16] and [21]) of the formalism is to construct such abstractions via induction. It does so using a formal definition of reality that makes as few assumptions as possible [6], in order to address symbol grounding [9] and other problems associated with dualism. In this context, cause and effect are statements as defined in 2. Returning to the example of Larry, instead of the variables A, C and R, we have an implementable language $\langle H, V, L \rangle$ and statements $c, r \in L$ such that if h is the current state of reality[8] then:

$$c \in h \leftrightarrow \textit{"Larry put on a raincoat"}$$
$$r \in h \leftrightarrow \textit{"It rained"}$$

As before, assume we have concluded $p(r \in h \mid c \in h) = 1$ from passive observation, the naive interpretation of which is that we can make it rain by forcing Larry to wear a coat. However, the statement associated with this intervention is not *just* $c =$ *"Larry put on a raincoat"* but a third $a \in L$ such that:

$$a \in h \leftrightarrow \textit{"Larry put on a raincoat because we forced him to"}$$



Because we're now dealing with statements, and because statements are sets of declarative programs which are inferred rather than given, we no longer need to explicitly define interventions in advance. Statements in an implementable language represent sensorimotor activity [16], and are formed via induction [6]. The observation of c is part of the sensorimotor activity a, meaning $c \subseteq a \subset h$ (if Larry is not wearing his raincoat, then it also cannot be true that we are forcing him to wear it). There is still no *do* operator, however $i = a - c$ may be understood as representing the identity of the party undertaking the intervention. If $i \neq \emptyset$ then it is at least possible to distinguish intervention from passive observation, in the event that a and c are relevant (we still need explain under what circumstances this is true). Whether intervention and observation are indistinguishable depends upon the vocabulary V, the choice of which determines if $i = \emptyset$, or $i \neq \emptyset$ (the latter meaning that it is distinguishable). Thus interventions are represented, but only to the extent that the vocabulary permits.

[8] Recall from definition 1 and 2 that a statement is true iff part of the present state.

Definition 9 (intervention). *If a is an intervention to force c, then $c \subseteq a$. Intervention is distinguishable from observation only where $c \subset a$.*

3.3 When will induction distinguish intervention from observation?

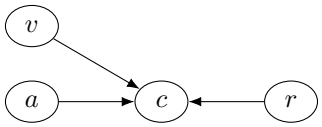
From **(prem. 1)** and **(prem. 2)** we have formal proof that choosing the weakest model maximises the probability of generalisation. There are many combinations of parent and child task for which generalisation from child to parent is only possible by selecting a model that correctly distinguishes the effects of intervention from passive observation (a trivial example might be a task informally defined as "predict the effect of this intervention"). It follows that to maximise the probability of generalisation in those circumstances the weakest model must distinguish between an intervention a and what it forces, c, so long as **(prem. 3)** is satisfied as described in definition 9, meaning $a \neq c$.

4 Awareness

We have described how an intervention a is represented as distinct from that which it forces, c. Induction will form models representing this distinction in tasks for which this aids completion. Now we go a step further.

4.1 Identity

Earlier we discussed $i = a - c$ as the identity of the party undertaking an intervention a. We might define a weaker identity as $k \subset i$, which is subset of any number of different interventions undertaken by a particular party. The *do* operator assumes the party undertaking interventions is given, and so we might think of k above as meaning "me". However, there is no reason to restrict emergent representations of intervention only to one's self. For example there may exist Harvey, who also intervenes to force c. It follows we may have v such that $c \subset v$, and v represents our observation of Harvey's intervention.



If $k \subseteq a - c$ can represent our identity as party undertaking interventions, it follows that $j \subseteq v - c$ may represent Harvey's. Both identities are to some extent context specific (an intervention undertaken elsewhere may produce something other than j for Harvey, or may be a subset of j), but these emergent identities nevertheless exist as a measurable quantity independent of the interventions with which they are associated.

Definition 10 (identity). *If a is an intervention to force c, then $k \subseteq a - c$ may function as an identity undertaking the intervention if $k \neq \emptyset$.*

One's own identity is used to distinguish interventions from passive experiences to facilitate accurate inductive inference in an interactive setting. It follows from **(prem. 1)** and **(prem. 2)** that every object that has an impact upon one's ability to complete tasks must *also* have an identity[9], because failing to account for the interventions of these objects would result in worse performance.

4.2 Intent

The formalism we are discussing originated as a theory of mind called "The Mirror Symbol Hypothesis" [21], and a mechanistic explanation of meaning in virtue of intent [16] (similar to Grice's foundational theory of meaning [22]). A statement is a set of declarative programs, and can be used as a goal constraint as is common in AI planning problems [23]. In the context of a task a model expresses such a goal constraint, albeit integrated with how that goal is to be satisfied [6]. If one is presented with several statements representing independent pairings of situation and decision (a task according to definition 3), then the weakest statement with which which one can derive the decisions from the situations (a model) is the *intent* those decisions served [16]. Thus, if identity k experiences interventions undertaken by identity j, then k can infer something of the intent of j by constructing a task definition and computing the weakest models [16]. This is a mechanistic explanation of how it is *possible* that one party may infer another's intent. Assuming induction takes place according to definition 8, then it is also *necessary* to the extent that k affect's j's ability to complete tasks. Otherwise, j's models would not account for j's interventions and so performance would be negatively impacted. However, a few interventions is not really much information to go on. Humans can construct elaborate rationales for behaviour given very little information, which suggests there is more to the puzzle. The Mirror Symbol Hypothesis argues that we fill in the gaps by projecting our own emergent symbols (models, in this context) representing overall, long term goals and understanding onto others in order to construct a rationale for their immediate behaviour [16], in order to empathise.

4.3 How might we represent The Mirror Symbol Hypothesis?

There would exist a task $\Omega = \langle S_\Omega, D_\Omega, M_\Omega \rangle$ which describes every decision k might ever make which meets some threshold of "good enough" [21, 16].

Definition 11 (Higher and lower level statements). *A statement $c \in L$ is higher level than $a \in L$ if $Z_a \subset Z_c$, which is written as $a \sqsubset c$*[10].

A model $m_\Omega \in M_\Omega$ is k's "highest level" intent or goal (given the threshold), meaning $Z_\Omega = D_\Omega$. Using m_Ω and k's observation of decision d made in situation s by j, k could construct a lower level model $m_\omega \sqsubset m_\Omega$ such that $d \in Z_s \cap Z_{m_\omega}$.

[9] Assuming interventions are distinguishable.

[10] Extension creates a lattice of statements, where the weakest statements represent the highest level of abstraction, and the strongest the lowest.

In other words, m_ω is a rationale constructed by k to explain j's intervention. A companion paper [24] on the computation of meaning explores this in more depth. For our purposes it suffices to say that in combining emergent causality and identity with The Mirror Symbol Hypothesis [21] and symbol emergence [16], we have a mechanistic explanation of the ability to reason about one's own identity and intent, and that of others. Following these arguments to their logical conclusion, the ability to predict how one's own intent is modelled by another is also of value in predicting that other's behaviour. This suggests that the maximisation of performance may necessitate identity k constructing a model of j's model of k, and $j's$ model of $k's$ model of j and so on to the extent permitted by the vocabulary (memory and other computational resources being a limitation).

4.4 Consciousness

We have described a means by which an agent may be, in a limited fashion, aware of itself, of others, of the intent of others and of the ability of others to model its own intent. By aware, we mean it has *access* to and will function according to this information (in the sense of access and functional consciousness). As Boltuc argues [11], phenomenal consciousness (characterised as first person functional consciousness) is already plausibly explained by the machine learning systems of today. Were we to integrate qualia into this system it might be as a function $q : L \to F$, where F is the set of all possible subjective feelings, and a statement $l \in L$ is sensorimotor activity.

4.5 Further discussion

An implementation of what we have described would construct an identity for anything and everything affecting its ability to complete tasks - even inanimate objects like tools, or features of the environment. Intent would be ascribed to those identities, to account for the effect those objects have upon one's ability to satisfy goals. Though this might seem a flaw, to do anything else would negatively affect performance. Interestingly, this is consistent with the human tendency [25] to anthropomorphise. We ascribe agency and intent to inanimate objects such as tools, the sea, mountains, the sun, large populations that share little in common, things that go bump in the night and so forth. Furthermore, the intent we ascribe tends to be a reflection our own understanding and experiences. This is consistent with The Mirror Symbol Hypothesis, which states that we ascribe intent by projecting our own overall models onto others [26].

References

[1] J. Pearl and D. Mackenzie. *The Book of Why: The New Science of Cause and Effect*. 1st. New York: Basic Books, Inc., 2018.

[2] P. A. Ortega et al. *Shaking the foundations: delusions in sequence models for interaction and control*. Deepmind, 2021.
[3] J. Pearl. "Causal Diagrams for Empirical Research". In: *Biometrika* 82.4 (1995), pp. 669–688. (Visited on 07/06/2022).
[4] J. Pearl. *Causality*. 2nd ed. United Kingdom: Cambridge Uni. Press, 2009.
[5] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Berlin, Heidelberg: Springer-Verlag, 2010.
[6] M. T. Bennett. *Enactivism & Objectively Optimal Super-Intelligence*. 2023.
[7] M. T. Bennett. *The Optimal Choice of Hypothesis Is the Weakest, Not the Shortest*. 2023.
[8] M. T. Bennett. *Computable Artificial General Intelligence*. 2022.
[9] S. Harnad. "The symbol grounding problem". In: *Physica D: Nonlinear Phenomena* 42.1 (1990), pp. 335–346.
[10] S. Franklin, B. J. Baars, and U. Ramamurthy. "A Phenomenally Conscious Robot?" In: *APA Newsletter on Philosophy and Computers* 1 (2008).
[11] P. Boltuc. "The Engineering Thesis in Machine Consciousness". In: *Techné: Research in Philosophy and Technology* 16.2 (2012), pp. 187–207.
[12] N. Block. "The Harder Problem of Consciousness". In: *Journal of Philosophy* 99.8 (2002), p. 391.
[13] D. Chalmers. "Facing Up to the Problem of Consciousness". In: *Journal of Consciousness Studies* 2.3 (1995), pp. 200–19.
[14] D. Ward, D. Silverman, and M. Villalobos. "Introduction: The Varieties of Enactivism". In: *Topoi* 36 (Apr. 2017).
[15] M. Wheeler. "Martin Heidegger". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Fall 2020. Stanford University, 2020.
[16] M. T. Bennett. "Symbol Emergence and the Solutions to Any Task". In: *Artificial General Intelligence*. Cham: Springer, 2022, pp. 30–40.
[17] M. T. Bennett and Y. Maruyama. *Intensional Artificial Intelligence: From Symbol Emergence to Explainable and Empathetic AI*. Manuscript, 2021.
[18] P. Bekinschtein et al. "A retrieval-specific mechanism of adaptive forgetting in the mammalian brain". In: *Nature Communications* 9.1 (2018), p. 4660.
[19] S. B. Berlin. "Dichotomous and Complex Thinking". In: *Social Service Review* 64.1 (1990), pp. 46–59.
[20] R. S. Nickerson. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises". In: *Review of General Psychology* 2.2 (1998), pp. 175–220.
[21] M. T. Bennett and Y. Maruyama. "Philosophical Specification of Empathetic Ethical Artificial Intelligence". In: *IEEE Transactions on Cognitive and Developmental Systems* 14.2 (2022), pp. 292–300.
[22] H. P. Grice. *Studies in the Way of Words*. Cambridge MA: Harvard University Press, 2007.
[23] H. Kautz and B. Selman. "Planning as satisfiability". In: *IN ECAI-92*. New York: Wiley, 1992, pp. 359–363.
[24] M. T. Bennett. *How to Compute Meaning & Lovecraftian Horrors*. 2023.

[25] E. G. Urquiza-Haas and K. Kotrschal. "The mind behind anthropomorphic thinking: attribution of mental states to other species". In: *Animal Behaviour* 109 (2015), pp. 167–176.

[26] M. T. Bennett. "Compression, The Fermi Paradox and Artificial Super-Intelligence". In: *Artificial General Intelligence*. Springer, 2022, pp. 41–44.