Emergent Causality and the Foundation of Consciousness

Michael Timothy Bennett[1]

[1]The Australian National University

April 17, 2024

Abstract

Awarded "Best Student Paper" and published in Proceedings of The 16th International Conference on Artificial General Intelligence, Stockholm, 2023.

To make accurate inferences in an interactive setting, an agent must not confuse passive observation of events with having intervened to cause them. The do operator formalises interventions so that we may reason about their effect. Yet there exist 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 a do operator, an intervention can be represented by a variable. We then argue that variables are abstractions, and that need to explicitly represent interventions in advance arises only because we presuppose these sorts of 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 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.

Emergent Causality and 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 intervened to cause them. The *do* operator formalises interventions so that we may reason about their effect. Yet there exist 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 a *do* operator, an intervention can be represented by a variable. We then argue that variables are abstractions, and that need to explicitly represent interventions in advance arises only because we presuppose these sorts of 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 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[1].

Keywords: causality · theory of mind · self aware AI · AGI

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 [2, 3]. Say we had two variables $R, C \in \{true, false\}$, where:

$$C = true \leftrightarrow \textit{“Larry put on a raincoat”} \text{ and } R = true \leftrightarrow \textit{“It rained”}$$

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 this is that we can make it rain

[1] Appendices are to be found on GitHub [1].

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. By intervening we are acting upon the system from the outside, to disconnect those factors influencing the choice of clothing. The "do" operator [4, 5] resolves this in that $do[C = true]$ represents the intervention. 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 [2]. 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 pareto optimal mathematical formalisms of general intelligence in an interactive setting which, given no explicit representation of intervention, make maximally accurate inferences [6, 7, 1]. 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 each intervention. Then, using one of the aforementioned formalisms, we argue that need to explicitly represent intervention as a variable only arises if we presuppose abstractions [8] like variables. If induction does not depend upon abstractions as given, then abstractions representing interventions 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 is a mechanistic explanation of awareness, in a narrow sense of the term. By narrow we mean functional, access, and phenomenal consciousness, and only if the latter is defined as "first person functional consciousness" [9, 10]; recognising phenomenal content such as light, sound and movement with one's body at the centre of it all [11]. To limit scope, we do not address "the hard problem" [12].

2 Additional background

This section introduces relevant background material. The reader may wish to skip ahead to section 3 and refer here as needed. In recognition of the philosophical nature of this topic 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, and in technical appendices available on GitHub [1]. To those more familiar with the agent environment paradigm, how exactly these definitions formalise cognition may seem unclear. Neither agent nor environment are defined. This is because it is a formalism of enactivism [13], 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 [7], which may be understood as context specific manifestations of intent, or snapshots of what bears some resemblance to "Being-in-the-world" as described by Heidegger [14]. In simpler terms, this reduces cognition to a finite set of decision problems [7]. 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 [15, p. 2].

2.1 List of definitions

Refer to the appendices [1] and the related papers [16, 17, 18] for further information regarding these definitions.

Definition 1 (environment).

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

Definition 2 (implementable language).

- *$\mathfrak{V} = \{V \subset P : V \text{ is } finite\}$ is a set whose elements we call **vocabularies**, one of which[2] we single out as **the vocabulary** $\mathfrak{v}$ for an implementable language.*
- *$L_{\mathfrak{v}} = \{l \subseteq \mathfrak{v} : \exists \phi \in \Phi \ (\forall p \in l : p(\phi) = true)\}$ is a set whose elements we call **statements**. $L_{\mathfrak{v}}$ follows from Φ and $\mathfrak{v}$. We call $L_{\mathfrak{v}}$ an **implementable language**.*
- *$l \in L_{\mathfrak{v}}$ is **true** iff the present state is ϕ and $\forall p \in l : p(\phi) = true$.*
- *The **extension of a statement** $a \in L_{\mathfrak{v}}$ is $Z_a = \{b \in L_{\mathfrak{v}} : a \subseteq b\}$.*
- *The **extension of a set of statements** $A \subseteq L_{\mathfrak{v}}$ is $Z_A = \bigcup_{a \in A} Z_a$.*

(Notation) *Z with a subscript is the extension of the subscript*[3].

Definition 3 ($\mathfrak{v}$-task). *For a chosen $\mathfrak{v}$, a task α is $\langle S_\alpha, D_\alpha, M_\alpha \rangle$ where:*

[2] The vocabulary $\mathfrak{v}$ we single out represents the sensorimotor circuitry with which an organism enacts cognition - their brain, body, local environment and so forth.

[3] e.g. Z_s is the extension of s.

- $S_\alpha \subset L_\mathfrak{v}$ *is a set whose elements we call* ***situations*** *of* α.
- S_α *has the extension* Z_{S_α}*, whose elements we call* ***decisions*** *of* α.
- $D_\alpha = \{z \in Z_{S_\alpha} : z \textit{ is correct}\}$ *is the set of all decisions which complete* α.
- $M_\alpha = \{l \in L_\mathfrak{v} : Z_{S_\alpha} \cap Z_l = D_\alpha\}$ *whose elements we call* ***models*** *of* α.

$\Gamma_\mathfrak{v}$ *is the set of all tasks for our chosen* $\mathfrak{v} \in \mathfrak{V}$.

(Notation) *If* $\omega \in \Gamma_\mathfrak{v}$*, then we will use subscript* ω *to signify parts of* ω*, meaning one should assume* $\omega = \langle S_\omega, D_\omega, M_\omega \rangle$ *even if that isn't written.*

(How a task is completed) *Assume we've a* $\mathfrak{v}$*-task* ω *and a hypothesis* $\boldsymbol{h} \in L_\mathfrak{v}$ *s.t.*

1. *we are presented with a situation* $s \in S_\omega$*, and*
2. *we must select a decision* $z \in Z_s \cap Z_{\boldsymbol{h}}$.
3. *If* $z \in D_\omega$*, then* z *is correct and the task is complete. This occurs if* $\boldsymbol{h} \in M_\omega$.

Definition 4 (probability). *We assume a uniform distribution over* $\Gamma_\mathfrak{v}$.

Definition 5 (generalisation). *A statement* l *generalises to* $\alpha \in \Gamma_\mathfrak{v}$ *iff* $l \in M_\alpha$*. We say* l *generalises from* α *to* $\mathfrak{v}$*-task* ω *if we first obtain* l *from* M_α *and then find it generalises to* ω.

Definition 6 (child and parent). *A* $\mathfrak{v}$*-task* α *is a child of* $\mathfrak{v}$*-task* ω *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.

Definition 7 (weakness). *The weakness of* $l \in L_\mathfrak{v}$ *is* $|Z_l|$.

Definition 8 (induction). α *and* ω *are* $\mathfrak{v}$*-tasks such that* $\alpha \sqsubset \omega$*. Assume we are given a proxy* $q_\mathfrak{v} \in Q$*, the complete definition of* α *and the knowledge that* $\alpha \sqsubset \omega$*. We are not given the definition of* ω*. The process of induction would proceed as follows:*

1. *Obtain a hypothesis by computing a model* $\mathbf{h} \in \arg\max_{m \in M_\alpha} q_\mathfrak{v}(m)$.
2. *If* $\mathbf{h} \in M_\omega$*, then we have generalised from* α *to* ω.

2.2 Premises

For the purpose of argument we will adopt the following premises:

> **(prem. 1)** To maximise the probability that induction generalises from α to ω, it is necessary and sufficient to maximise weakness. [1]

For our argument this optimality is less important than the representation of interventions it implies. In any case the utility of weakness as a proxy is not limited to lossless representations or optimal performance. Approximation may be achieved by selectively forgetting outliers[4], a parallel to how selective amnesia

[4] 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.

[19] can help humans reduce the world to simple dichotomies [20] or confirm preconceptions [21]. Likewise, a task expresses a threshold beyond which decisions are "good enough" [22]. The proof of optimality merely establishes the upper bound for generalisation. As a second premise, we shall require the emergence or presupposition of representations of interventions:

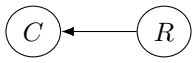
> **(prem. 2)** To make accurate inductive inferences in an interactive setting, an agent must not confuse the passive observation of an event with having intervened to cause that event. [2]

3 Emergent Causality

The formalism does not presuppose an operator representing intervention. Given our premises, we must conclude from this that either that **(prem. 1)** is false, or induction as in definition 8 will distinguish passive observation of an event from having intervened to cause that event.

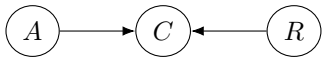
3.1 The *do* operator as a variable in disguise

In the introduction we discussed an example involving binary variables R (rain) and C (raincoat). From $p(R = true \mid C = true) = 1$ we drew the absurd conclusion that if we intervene to make $C = true$, we can make it rain. The true relationship between R and C is explained by 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:

By intervening in the system, we are acting upon it from the outside. In doing so we disconnect those factors influencing the choice of clothing. The *do* operator lets us express this external influence. However, if we don't have a *do* operator there remains another option. Interventions can be represented by additional variables [23][5], so that we are no longer intervening in the system from outside. For example $do[C = true]$ might be represented by A such that $p(C = true \mid A = true) = 1$ and $p(C \mid A = false) = p(C)$:



We can now represent that $p(R = true \mid C = true, A = true) = p(R = true) \neq p(R = true \mid C = true, A = false) = 1$. This expands the system to include an action by a specific actor, rather than accounting for interventions originating outside the system (as the *do* operator does).

[5] This preprint has been corrected post-publication to include this citation of Dawid, as we were previously unaware of his work.

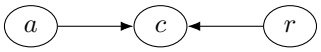
3.2 Emergent representation of interventions

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 [7] and [22]) of the formalism is to construct such abstractions via induction. It does so by formally defining reality (environment and cognition within that) using as few assumptions as possible [1], in order to address symbol grounding [8] 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 variables A, C and R we have a vocabulary $\mathfrak{v}$, and $c, r \in L_{\mathfrak{v}}$ which have a truth value in accordance with definition 2:

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

As before, assume we have concluded $p(r \mid c) = 1$ 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:

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



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, and are formed via induction [7, 1]. The observation of c is part of the sensorimotor activity a, meaning $c \subseteq a$ (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.

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

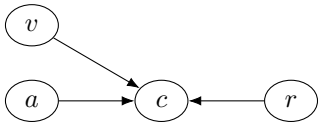
3.3 When will induction distinguish intervention from observation?

From **(prem. 1)** we have 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. 2)** is satisfied as in def. 9, s.t. $a \neq c$.

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. 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 (another intervention may produce something other than j, or a subset of j, for Harvey), but these emergent identities still exist as a measurable quantity independent of the interventions with which they're 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)** that every object that has an impact upon one's ability to complete tasks must *also* have an identity[6], because failing to account for the interventions of these objects would result in worse performance.

4.1 Intent

The formalism we are discussing originated as a mechanistic explanation of theory of mind called "The Mirror Symbol Hypothesis" [22], and of meaning in

[6] Assuming interventions are distinguishable.

virtue of intent [7] (similar to Grice's foundational theory of meaning [24]). A statement is a set of declarative programs, and can be used as a goal constraint as is common in AI planning problems [25]. In the context of a task a model expresses such a goal constraint, albeit integrated with how that goal is to be satisfied [7, 1]. If one is presented with several statements representing decisions, and the situations in which they were made (a task according to definition 3), then the weakest statement with which which one can derive the decisions from the situations (a model) is arguably the *intent* those decisions served [7]. 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 [7]. 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 (either tasks or models, in this context) representing overall, long term goals and understanding onto others in order to construct a rationale for their immediate behaviour [7], in order to empathise.

4.2 How might we represent The Mirror Symbol Hypothesis?

Assume there exists a task Ω which describes every decision k might ever make which meets some threshold of "good enough" [22, 7] at a given point in time.

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 $a \sqsubset c$.*

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 (the observation of which would also be a decision), k could construct a lower level model $m_\omega \sqsubset m_\Omega$ such that $d \in Z_s \cap Z_{m_\omega}$. In other words, m_ω is a rationale constructed by k to explain j's intervention. Related work explores this in more depth [7, 22]. For our purposes it suffices to point out that in combining emergent causality, identity, The Mirror Symbol Hypothesis [22] and symbol emergence [7], we have a mechanistic explanation of the ability to reason about one's own identity and intent, and that of others, in terms of interventions. Likewise the ability to predict how one's own intent is modelled by another is also of value in predicting that other's behaviour. In tasks of the sort encountered by living organisms, optimal performance would 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 greatest extent permitted by $\mathfrak{v}$ (the finite memory and any other limitations one's ability to represent predictions of predictions of predictions ad infinitum).

4.3 Consciousness

We have described a means by which an agent may be 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 (access and functional consciousness, contextualising everything in terms of identities and their intent). Boltuc argues that phenomenal consciousness (characterised as first person functional consciousness) is explained by today's machine learning systems [10]. We would suggest his argument extends to our formalism, and in any case if qualia are a mechanistic phenomenon then they are already represented by the vocabulary of the implementable language. What is novel in our formalism is not just that it points out that causal inference may construct identity and awareness, but that it does so with a formulation that also addresses enactive cognition, symbol emergence and empathy [22, 7].

Anthropomorphism: 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 [26] 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.

Fragmented identities: It is also interesting to consider what this says of systems which are less than optimal (do not identify the weakest hypothesis), or which do not use a vocabulary which permits the construction of one identity shared by all of the interventions it undertakes. Such a thing might construct multiple unconnected identities for itself, and ascribe different intentions to each one. Likewise if the model constructs multiple identities for what is in fact the same object, it may hallucinate and hold contradictory beliefs about that object.

References

[1] M. T. Bennett. *Appendices*. Version 1.2.1. 2023. DOI: 10.5281/zenodo.7641742. URL: github.com/ViscousLemming/Technical-Appendices.

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

[3] P. A. Ortega et al. *Shaking the foundations: delusions in sequence models for interaction and control*. Deepmind, 2021.

[4] J. Pearl. "Causal Diagrams for Empirical Research". In: *Biometrika* 82.4 (1995), pp. 669–688. (Visited on 07/06/2022).

[5] J. Pearl. *Causality*. 2nd ed. United Kingdom: Cambridge Uni. Press, 2009.

[6] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Berlin, Heidelberg: Springer-Verlag, 2010.
[7] M. T. Bennett. "Symbol Emergence and the Solutions to Any Task". In: *Artificial General Intelligence*. Cham: Springer, 2022, pp. 30–40.
[8] S. Harnad. "The symbol grounding problem". In: *Physica D: Nonlinear Phenomena* 42.1 (1990), pp. 335–346.
[9] S. Franklin, B. J. Baars, and U. Ramamurthy. "A Phenomenally Conscious Robot?" In: *APA Newsletter on Philosophy and Computers* 1 (2008).
[10] P. Boltuc. "The Engineering Thesis in Machine Consciousness". In: *Techné: Research in Philosophy and Technology* 16.2 (2012), pp. 187–207.
[11] N. Block. "The Harder Problem of Consciousness". In: *Journal of Philosophy* 99.8 (2002), p. 391.
[12] D. Chalmers. "Facing Up to the Problem of Consciousness". In: *Journal of Consciousness Studies* 2.3 (1995), pp. 200–19.
[13] D. Ward, D. Silverman, and M. Villalobos. "Introduction: The Varieties of Enactivism". In: *Topoi* 36 (Apr. 2017).
[14] M. Wheeler. "Martin Heidegger". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Fall 2020. Stanford University, 2020.
[15] M. T. Bennett and Y. Maruyama. *Intensional Artificial Intelligence: From Symbol Emergence to Explainable and Empathetic AI*. Manuscript, 2021.
[16] M. T. Bennett. *Computational Dualism and Objective Superintelligence*. 2023. URL: `arxiv.org/abs/2302.00843`.
[17] M. T. Bennett. "The Optimal Choice of Hypothesis Is the Weakest, Not the Shortest". In: *Artificial General Intelligence*. Springer, 2023, pp. 42–51.
[18] M. T. Bennett. "On the Computation of Meaning, Language Models and Incomprehensible Horrors". In: *Artificial General Intelligence*. Springer, 2023, pp. 32–41.
[19] P. Bekinschtein et al. "A retrieval-specific mechanism of adaptive forgetting in the mammalian brain". In: *Nature Communications* 9.1 (2018), p. 4660.
[20] S. B. Berlin. "Dichotomous and Complex Thinking". In: *Social Service Review* 64.1 (1990), pp. 46–59.
[21] R. S. Nickerson. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises". In: *Review of General Psychology* 2.2 (1998), pp. 175–220.
[22] 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.
[23] A. P. Dawid. "Influence Diagrams for Causal Modelling and Inference". In: *International Statistical Review / Revue Internationale de Statistique* 70.2 (2002), pp. 161–189. ISSN: 03067734, 17515823. URL: `http://www.jstor.org/stable/1403901` (visited on 02/22/2024).
[24] H. P. Grice. *Studies in the Way of Words*. Cambridge MA: Harvard University Press, 2007.
[25] H. Kautz and B. Selman. "Planning as satisfiability". In: *IN ECAI-92*. New York: Wiley, 1992, pp. 359–363.

[26] 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.