Contents

Research Articles

Book Review

Organon F 32 (4) 2025: 380–401 ISSN 2585-7150 (online)
https://doi.org/10.31577/orgf.2025.32401

RESEARCH ARTICLE

The Dual Nature of Explicability in AI Ethics

Hyundeuk Cheon*

Received: 2 October 2024 / Revised: 9 April 2025 / Accepted: 13 April 2025

Abstract: Despite the significance of explicability in AI ethics, the principle of explicability remains subject to several unresolved issues, including its moral status, purpose, and the recipients of explanations. First, this paper proposes treating explicability as a prima facie duty to make machine learning algorithms explicable. Second, the dual nature of explicability is highlighted. It is claimed that explicability is for the warranted trust of decision-recipients in the algorithmic decisions as well as for enhancing the autonomy of decision-makers.

Keywords: Algorithm; explicability; explainability; trust; trustworthiness; autonomy.

1. A Call for the Principle of Explicability

Machine learning algorithms (hereafter, ML algorithms) are increasingly being utilized in complex, real-world decision-making, which is often of ethical significance. While the obvious cases are autonomous weapon systems and self-driving cars, these algorithms are also used in ordinary decision-making, from reviewing job applications, assessing loan applications, and approving parole to predictive policing. They are designed to help allocate

* Seoul National University
https://orcid.org/0000-0002-7569-7776
✎ Seoul National University, Department of Science Studies, Building 25-419, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, Korea, 08826
✉ hdcheon@snu.ac.kr

social services, decide on promotions or terminations, determine credit scores, or estimate a person's risk of committing crimes. With rapid technological development, there are growing concerns about the ethical and responsible uses of AI algorithms. In particular, the demand for explainability or transparency in AI ethics has attracted considerable attention. For example, European Union's *General Data Protection Regulation* is often interpreted as incorporating the 'right to explanation' when someone is affected by automated decision-making (Goodman and Flaxman, 2017). The Future of Life Institute (2017) declared the Asilomar AI principles, which emphasized the transparency of algorithms when it causes harm or is involved in judicial decision-making.[1] Microsoft's CEO Satya Nadella (2016) also calls for a similar requirement of intelligibility in terms of understanding "how the technology works and what its rules are."[2]

Luciano Floridi, one of the leading figures in AI ethics, suggested that similar principles - the principles of explainability, transparency, interpretability, and accountability - can be incorporated under an overarching principle, what he called the principle of explicability (Floridi et al., 2018; Floridi and Cowls, 2019). Floridi and his colleagues put forth the notion of explicability as encompassing both the epistemological sense of intelligibility and the ethical sense of accountability. The former concerns an answer to the question 'How does it work?' and the latter concerns an answer to the question 'Who is responsible for the way it works?' As long as the notion of explicability is general and robust, there is a wide consensus on the significance of the explicability of ML algorithms.

[1] It includes two kinds of transparency: the failure and the judicial transparency. According to the failure transparency principle, if an AI system causes harm, it should be possible to ascertain why. According to the Judicial transparency, any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority (AI principle 2017).

[2] "We should be aware of how the technology works and what its rules are. We want not just intelligent machines but intelligible machines. Not artificial intelligence but symbiotic intelligence. The tech will know things about humans, but the humans must know about the machines. People should have an understanding of how the technology sees and analyzes the world. Ethics and design and in hand" (Nadella 2016).

However, there are several unresolved issues on the principle of explicability: its moral status, what it is for, to whom AI needs to be explicable, and what kinds of explanation should be given to meet the principle. First, the normative status of the principle is not clearly stated (e.g., whether it is a categorical requirement or a recommendation). Second, regarding why the principle is needed, there is a controversy between the trust-based view and the autonomy-based view. While some argue that explicable AI is for human autonomy, many scholars claim that it is required to enhance users' trust in the system and its generated results. Third, there is no agreement on whether explicability ought to be directed toward decision-makers using the algorithms or toward decision-recipients affected by the algorithm-assisted decisions. Further, it is required to provide the criteria on which we assess the intelligibility and accountability of AI. Although it is sometimes recognized that different explanations need to be given to different stakeholders for different purposes, the problem of what kinds of explication are required (and how they meet the criteria) has yet to be thoroughly investigated.[3] For the reason of space, however, we confine our discussion to what-, why-, and whom-questions.

This paper aims to contribute to explicating the concept of explicability rather than providing a survey or description of its current usage in the relevant literature. By "explication," as defined in Carnap (1950, 1955), I mean a prescription of how we should characterize the concept in order to use it more productively. It involves replacing the unclear concept (the explicandum) with a clearer one (the explicatum) to serve the goals of the concept better. Our primary focus is to demonstrate the dual nature of explicability by addressing the why- and whom-questions. While this paper does not present entirely novel arguments for particular views over others, it adopts a methodological strategy of dissecting and rearranging the existing materials from the current literature on explicability. I contend that the seemingly plausible arguments failed to support the claim they are supposed to because they are ignorant of the dual nature of explicability. Some arguments are cogent in demonstrating one aspect of explicability but fail to account for another aspect, and *vice versa.* To a coherent set of answers to

[3] The question of what kinds of explanations are needed for different contexts will be addressed in a separate article.

the why and whom-questions, it is crucial to take into account the duality of explicability.

This paper will proceed as follows. In Section 2, I begin with setting the stage by examining how data science usually works in practice and suggest the principle of explicability as a prima facie duty to make ML algorithms explicable. In Section 3, I clarify the who(m)-question, addressing who is responsible for making algorithms explicable and to whom the explanation is owed. In Section 4, I delve into reasons for demanding explicability, highlighting the duality of autonomy and trustworthiness. In Section 5, it is claimed that the explication should be directed toward decision-makers as well as decision-recipients.

2. Setting the Stage

In our context under consideration, the primary function of ML algorithms is to offer prediction-based classification or ranking for specific purposes. It is often claimed, for example, that ML algorithms estimate the probability of a prisoner committing crimes again to determine their eligibility for release on parole or predict an applicant's credit score to classify her as qualified or unqualified for a bank loan (Pasquale, 2015; O'Neil, 2016). Of course, before the advent of ML algorithms, individuals have routinely been subject to classification: people have been ranked on credit score, deemed as qualified or disqualified for insurance, or accepted or rejected for various applications. Now, these classificatory tasks are being automated with ML algorithms.

How do automated or algorithm-assisted decision-making systems work in real-world cases? Roughly speaking, the decision-making processes consist of several steps that need to be iterated: defining the problem, data collection, building algorithm models, training and testing models with collected data, and application in real-world situations.[4] Think of a loan application system as an instance. First, the problem should be clearly defined:

[4] Roose (2023) introduces similar steps in explaining how AI technology actually works: 1) set a goal, 2) collect lots of data, 3) build your neural network, 4) train your neural network, 5) fine-tune your model, and 6) launch carefully.

classifying an applicant based on her credit and ability to repay the debt. Then, developers build models and collect data to train and test the model. If the tested model turns out to be successful, it will be applied in the decision-making of loan applications. Such algorithmic decision-making systems are to be used in a wide range of real-world situations.

Let me, in this paper, focus on situations where the use of ML algorithms is morally significant in that they involve direct assessment of personal capacities, characters, or properties that would affect the condition or quality of their lives (i.e., judicial sentencing, job interview, or predictive policing).[5] Given that the algorithms are used in this context, precisely what does the principle of explicability state? Although Floridi and colleagues (Floridi et al., 2018; Floridi and Cowls, 2019) put forth explicability as an overarching principle, they are not explicit on the normative status of the principle. As an ethical principle, it might be understood as a categorical imperative that ML algorithms be explicable or a recommendation that explicable algorithms be more desirable than inexplicable ones. Or, the principle might be interpreted as saying that it is morally wrong to use inexplicable AI in specific contexts.[6]

To clarify the normative status of the principle, we can take advantage of an analogy with biomedical ethics. In the meta-analysis of AI ethics guidelines, Floridi and Cowls (2019) found that AI ethics share four core principles of bioethics: beneficence, non-maleficence, autonomy, and justice

[5] I do not intend that those situations mentioned exhaust all the cases where using ML algorithms is ethically significant. Further, there are many other situations where algorithms have nothing to do with assessing personal capacities, characters, or properties (e.g., voice assistants), or no one is affected significantly by the algorithms in a moral sense (e.g., spam filters). However, the issue of accountability, the ethical aspect of explicability, arises when someone is affected by algorithm-assisted decision making.

[6] The failure transparency of Asilomar AI principles (2019) states that if an AI system causes harm, it should be possible to ascertain why. It sounds like a categorical imperative which should be obeyed when there is harm caused by AI systems. London (2019) takes the principle as a recommendation to prioritize explainability over predictive accuracy and criticizes it as unwarranted in the healthcare field.

(Beauchamp and Childress, 2012).[7] It is worth noticing that the four principles are usually regarded as prima facie duties. As they are not exceptionless requirements or imperatives, each principle needs to be weighed against other principles when applied to a concrete case and, when warranted, can be overruled by other principles. While they are instrumental in reflecting on moral problems and working toward an ethical resolution, they do not provide easy, ready-made solutions to particular cases. Likewise, we can take the principle of explicability as a prima facie duty for contemplating the moral problems of algorithmic decision-making systems, not merely an unqualified requirement for every application of AI algorithms. More specifically, I suggest the explicability principle as involving a prima facie duty to make ML algorithms explicable when applied in morally significant situations. As a prima facie duty, explicability does not necessarily take priority over others.[8] Nonetheless, the principle can be overruled only when there are good reasons justifying doing so.

Efforts to develop explicable ML systems can create tensions with other important values, including predictive accuracy, efficiency, privacy, and fairness. For the ethical development and deployment of ML systems, it is necessary to carefully manage the trade-offs among these competing values. This requires consideration of the specific context of deployment since no single solution is suitable for every situation. Suppose the ML algorithms are used in medical diagnosis, which often demonstrate superior predictive accuracy despite their highly complex and inscrutable structures. In this medical context, patient well-being is at stake, and diagnostic errors have severe consequences. Consequently, prioritizing accuracy over explicability can be ethically justified. Another type of trade-off involves the tension between explicability and efficiency. In situations demanding urgent decision-making (e.g., real-time financial trading or emergency response systems), rapid action is vital. In such demanding scenarios, the time delays

[7] It does not mean that they do offer a perfect translation. It only means that various ethical principles in different guidelines can be incorporated into four overarching principles with slightly different connotations.

[8] Of course, the remaining question is how to deal with a situation where different principles conflict. That is a subject to further discussion, depending on the context.

caused by generating detailed explanations may be practically untenable. Hence, in these contexts, operational efficiency can override the requirement for transparency.

Additional trade-offs become apparent when it comes to considering data privacy and fairness. For example, when the automated system is used to assess job applicants, providing exhaustive explanations might inadvertently compromise the applicant's privacy by revealing sensitive personal information. Moreover, full transparency can sometimes unintentionally undermine fairness. This can occur when simple, interpretable models rely on proxy variables closely correlated with protected characteristics (e.g., race or gender). Then, the principle of explicability may conflict with the demand for anti-discrimination. Paradoxically, a reduced degree of explicability in such situations might help to preserve privacy and promote fairness. Determining when the principle of explicability can be overridden requires careful, context-specific ethical deliberation. Decision-makers must assess the potential benefits we can get by improved accuracy or efficiency and the risks posed by diminishing privacy or fairness. Critically, the ethical legitimacy of any trade-off depends upon the acceptance or consent of the stakeholders affected by the decision-making systems.

Given that explicability is a prima facie duty, we must ask who holds the duty, to whom, and why.

3. Clarifying the Who and Whom Questions

Two questions, 'Why does the principle of explicability matter?' and 'Who has the duty to make algorithms explicable to whom?' are intimately related. As demonstrated subsequently, the answer to the first question informs the latter. In order to clarify the who(m)-questions, it is crucial to identify the key stakeholders involved in using automated decision-making systems. Among many stakeholders, five roles participate primarily in developing and utilizing ML algorithms: algorithm-developers, algorithm-users (and decision-makers), managing-users, decision-recipients, and regulators (cf. Arrieta et al., 2020; Langer et al., 2021). It appears that developers, algorithm-users, and decision-recipients are the most directly involved stakeholders with respect to the who/whom questions.

Algorithm-developers are those who are engaged in designing and developing ML algorithms. The developed and tested algorithms are used by algorithm-users. By algorithm-users, I mean those who make decisions using ML algorithms. Usually, algorithm-users are professionals (e.g., judges, police officers, and human resources managers) working in tandem with a system, or those who use the system make decisions that affect individuals. Decision-recipients (affected parties) are individuals or groups of individuals who are actually or potentially influenced by the system's decision. Typically, the three roles separately hold. Designers build the system on users' requests, and users make decisions about the applicants who are affected by the decision. For example, when algorithms are used in reviewing loan applications, developers design and develop the review algorithm, bank officers use it to judge whether applicants are eligible, and the decision made influences applicants.

In addition to the three main roles, other stakeholders are indirectly but significantly engaged in decision-making processes: managing-users and regulators. Managing-users (e.g., company, committee, etc.) decide to adopt an automated decision-making system with which algorithm users make decisions. Although they are not working with particular ML algorithms, they request developers to build a decision-making system with the desired specification and supervise the uses of the system. Thus, managing users are users of algorithms in an indirect sense. Finally, there are regulators (e.g., the government) who monitor and regulate the whole process and intervene in the cases where necessary. They are expected to establish moral and legal standards for the general use and development of automated systems. This class of stakeholders plays a unique role since they operate as a "watchdog" for both the systems themselves and the way in which they interact with the other stakeholders.

They are five different roles that people can play in decision-making processes. I prefer to talk about the roles rather than types of stakeholders because a single stakeholder can often play multiple roles simultaneously. For example, a CEO of an IT company who actively contribute to developing an algorithm for reviewing job applicants may also use it to make decisions, thus playing the roles of a developer, algorithm-user, and managing-user. Alternatively, one person (or a group of people) can be both a decision-maker and decision-recipient, where their decisions affect themselves. When we use a recommendation system for choosing books or movies (on Amazon or Netflix), we play the roles of algorithm-users and decision-recipients in that we make decisions affecting ourselves. In these cases, people decide whether to accept or reject the algorithm's suggestions or to use them for their actions. Still, the distinction of roles between designers, decision-makers, and decision-recipients holds.[9] Further, in our context of interest, where the issue of accountability is salient, the asymmetry between algorithm-users who make decisions with ML algorithms and decision-recipients who are affected by the decisions is assumed.

Given that the principle of explicability is about a prima facie duty to make ML algorithms explicable, we need to distinguish two who-questions: "Who has the duty?" and "To whom should the algorithms be explicable?" Let us call the former the who-question and the latter the whom-question. The answer to the former, I believe, is not very controversial. Anyone involved in designing, building, and using ML algorithms in ethically significant situations ought to contribute to making them explicable. First of all, algorithm-developers have a duty because they are the only ones who can make algorithms explicable (or inexplicable) in a direct sense. Nevertheless, the duty must be distributed among managing-users and regulators because they are supposed to contribute to developing and using the algorithms. Managing-users ought to have developers make ML algorithms explicable, while regulators must guarantee that the algorithms used in morally significant situations are explicable.

If we have a consensus on who has the duty, then the remaining question is: ML algorithms should be explicable to whom? One obvious response would be that ML algorithms should be explicable to every role of

[9] Medical AIs (e.g., imaging) are located in more complex situations. Developers and engineers build the medical AI system at the request of health professionals, who make use of the system to diagnose and treat diseases. Interestingly, decision-making in AI-assisted medicine is distributed among doctors and patients, although the final decision should be made by patients. With the aid of the ML algorithm, health professionals determine what the disease is and suggest a promising treatment. Patients accept (or reject) the treatment based on the doctor's suggestion (based on ML) and their own values and preferences.

stakeholders, including developers and regulators. Developers might think explicable AI is desirable because it helps to examine the limitations and errors and improve the performance of a system (e.g., debugging). To build a reliable system, however, it is unnecessary to make it explicable. Of course, there is a trivial sense that ML algorithms must be explicable to developers if they ought to be explicable to *anyone*. For the developers are those who are able to make it explicable and provide the explication to other stakeholders. From the regulator's perspective, explicability is needed to supervise and regulate the use of algorithms in morally significant decision-making. It does not necessarily mean the algorithms should be explicable to regulators in every morally loaded case. Instead, regulators might request the explicable AIs on behalf of decision-recipients and ordinary citizens because they are actual or potential parties concerned who are directly assessed and affected by the algorithmic decision-making systems.

The whom-question we are asking is more specific: to whom should the explanation be intelligibly given when ML algorithms are used in morally significant cases? According to a dominant view, what I call "the recipient-oriented view," algorithms should be explicable to individuals affected by the decision-making system's outputs (e.g., Grote and Berens, 2020; Kim and Routledge, 2021; Watcher et al., 2018). In other words, the developers ought to make the algorithms explicable so that the decision-makers using the algorithms can provide the decision-recipients with an explanation (e.g., "how the decision was produced"). Although most of the literature focuses on the explicability toward the decision-recipients, there is an alternative view, what we might call "the user-oriented view." For example, Robbins (2019) claims that explicability is not directed toward the person subject to the algorithm's decision, and algorithms should be explicable to those who make a decision using the algorithms. In other words, the algorithm-designers ought to make it explicable to decision-makers.

I suggest that the recipient-oriented view and the user-oriented view are not mutually exclusive but complementary to each other. In this regard, the problem with the two views lies in the assumption that explicability is directed to only one class of stakeholders involved in the decision-making processes, which has not been justified. The recipients-oriented view mistakenly assumes that the decision-recipients are those only who have the right to receive the explanation of the algorithm's decision. Robbins rightly highlights the significance of being explicable to decision-makers, which has been largely ignored by the proponents of the recipient-oriented view. However, he is also mistaken to maintain that explication is not directed toward the decision-recipients.[10] It is worth noticing that different stakeholders might want explicable AI for different purposes. An algorithm explicable to one stakeholder might be inexplicable to other stakeholders. This consideration motivates us to consider the possibility that while each view reflects a part of the whole process, explicability should be given to both decision-makers and decision-recipients. The duality of explicability with respect to the whom-question is intimately related to the dual goals of explicability. As the answer to the whom-question is informed by the way we answer the why question, we will discuss why the explicability of algorithms matters in the next section.

4. Why Algorithms Should Be Explicable

(1) Opacity and Request for Explicability

What is explicability for? Why should we adopt the explicability principle alongside other basic principles?[11] Regarding these questions, two main views comprise most of the literature: the trust-based and the autonomy-based views. Some scholars argue for the autonomy-based view, according to which the explicability helps humans working with algorithm models make their own informed decisions (e.g., Robbins, 2019). Others have advocated the so-called "trust-based view," according to which the explicability

[10] Robbins claims that "the person using the algorithm is the person that the explanation should be directed towards—not the person subject to the decision of the algorithm"(2019, p. 503). The reason why the claim is not justified will be examined in Section 6.

[11] From the managing user's perspective, it is necessary to comply with relevant regulations (e.g., GDPR's a right to explanation). However, such a compliance is merely a derivative reason while we are seeking for answers in a more fundamental level.

of algorithms is instrumental in enhancing the trust of decision-recipients in ML algorithms (e.g., Kim and Routledge, 2021).

Before we go further, the clarification of the trust-based view is in order. It is crucial to distinguish between trust and trustworthiness (Simon, 2013; McLeod, 2021). Note that people can trust someone who is untrustworthy and sometimes do not trust trustworthy persons. That is, the act of trust of the trustor in the trustee should not be conflated with the trustworthiness as a property possessed by the trustee. For business purposes, it would be advantageous to enhance people's trust in AI systems; however, the increase in trust itself is not constitutive of a moral obligation to make the system explicable. Instead, one can maintain that developers are morally obliged to make trustworthy algorithms[12], which requires explicability. In other words, we want people's trust in algorithms to be justified or warranted. In the following, when I refer to the trust-based view, I will mean the idea that explicability is required to ensure that decision-recipients' trust in algorithms is warranted. In the remainder of this section, I claim that the two views - the autonomy-based and the trust-based view - are not competing but complementary to each other.

The request for explicable AI seems to stem from the opaque nature of ML algorithms. In contrast to GOFAI, recent ML algorithms (i.e., deep neural net) are not transparent (for the forms of opacity, see Burrell, 2016).[13] It is widely held that the black-box nature of algorithms calls for

[12] The call for trustworthy AI abounds. For instance, European Commission (2019) has presented *Ethics Guidelines for Trustworthy AI* and OECD (2019) published *Recommendation on AI* which emphasize the "international co-operation for trustworthy AI."

[13] There are different senses in which ML algorithms are not transparent. Basically, the opacity of ML algorithms means that the recipients of algorithm's output have no understanding of why the decision-output has been made (and the inputs themselves are unknown or partially known). The opacity might mean intentional secrecy (no information open to recipients) or technical illiteracy (not easily understandable to non-experts). But there is a sense in which ML algorithms are fundamentally opaque even to engineers, which means "the mismatch between mathematical optimization in high-dimensionality character of ML and semantic interpretation demanded for human intelligibility" (Burrell, 2016). The fundamental sense of opacity is our focus in this paper.

the principle of explicability when used in morally loaded cases. If the outcomes of the black boxes have no moral significance, then explicability is not demanded.[14] For instance, no one would argue for a duty to make AlphaGo (Go-playing algorithm) explicable. Hence, explicability is called for when black boxes are used in morally significant situations. Our question is why it is so.

I claim that explicability is called for when opaque technologies threaten autonomy or trustworthiness. To put it in another way, the demand for explicability does not arise when 1) the technologies can be justifiably trusted and 2) users are able to control the uses of the technologies in that they judge whether (and/or when) they can use it or not based on the knowledge of its capabilities and limitations. For example, think of pills whose physiological mechanism on how it works is unknown. Still, if they are proven safe and effective, and we know when we can use them (and when we should not), there is no need to explain how it works. The problem is that the opaque nature of ML algorithms can threaten trustworthiness and autonomy (control to decide).

(2) Undermined Trustworthiness Calls for Explicability

Opacity undermines trustworthiness or warranted trust. Opaque algorithms make it difficult to judge whether decision-recipients' trust in them is warranted. Some scholars think that reliability is better for conferring trust than explanation.[15] For example, London (2019) argues that prioritizing explicability over reliability is misleading and criticizes the call for an explanation as unwarranted. Indeed, there are many cases where new

[14] Robbins (2019) made a similar point.

[15] Even when ML algorithms have been shown to increase the accuracy of diagnosis, a lack of explanation for the diagnosis can lower doctors' trust in the algorithms (Ribeiro, Singh, and Guestrin, 2016; Creel, 2020). However, it is unclear whether people will trust AI more when given some explanation. Recently, there is a growing body of empirical work on how providing an explanation affects people's trust in ML and its decision (e.g., Lu et al., 2019). But they usually focus on types of explanation (i.e., decision tree or diagram) to effectively raise people's trust. Here we are not concerned with mere increase of trust but with the warrant of the trust or trustworthiness.

technology can be trusted when it works well by reliably generating the desired output. Pills might be justifiably 'trusted' if their functionality is well-proven. Thus, we can ask whether explicability is required when ML algorithms turn out to be reliable. London and others would say that reliability suffices while explicability is not required.

It is widely held that we should (or do) pursue AI technologies that are not merely reliable but also trustworthy (European Commission, 2019; OECD, 2019). One of the consensuses of philosophical literature on trust and trustworthiness is that trust is not mere reliance, although it is a kind of reliance. As Baier (1986, p. 235) succinctly put it, "trusting can be betrayed, or at least let down, and not just disappointed." While reliance or reliability can be defined in terms of rational expectation, trust or trustworthiness is often defined in terms of normative or moral expectations of trustors. Unlike mere reliance, trust includes a normative expectation whose violations induce to betrayal, not merely disappointment. Trustworthiness can be defined by the property of a person who can be justifiably trusted. As trust can be betrayed, it is valuable only when directed to trustworthy persons.

To establish trustworthiness, two requirements must be met: competency and responsiveness. As articulated by Jones (2012), a person is trustworthy with respect to the trustor in domain of interaction D if and only if "she is competent with respect to the domain, and she would take the fact that [the trustor] is counting on her (...) to be a compelling reason for acting as counted on" (pp. 70–71). Competency refers to a person's ability to fulfill the trustor's expectation with respect to the tasks she is supposed to do. Responsiveness, on the other hand, involves taking into account and responding to the trustor's interests and values. While a technological artifact can be deemed competent if it can fulfill the user's expectations regarding its task, it cannot meet the responsiveness condition.

To apply the notion of trust and trustworthiness to AI technologies, we need to consider a socio-technical system incorporating technological artifacts, human agents, and institutional structure (Nickel et al., 2010, Rieder et al., 2021). Under this interpretation, the trustee is not a technological artifact but a network of technical objects and human agents. Accordingly, when someone 'trusts' a particular technology, she trusts a socio-technical system that includes designers, operators, and other stakeholders interacting within a regulatory or legal framework. In this regard, pills or other technical products (as a component of socio-technical systems) can be regarded as trustworthy to the extent that the human agents involved in developing and operating them are trustworthy, and the evaluative system for testing them is validated (e.g., industrial standards or certificates). It is the socio-technological system that can be responsive to the user's interests and values.

It is worth noticing that reliability in the sense of well-functioning does not guarantee the trustworthiness of technology in cases where due process is crucial. For an ML algorithm to be trustworthy, it must be both competent and responsive. ML algorithms count as competent if they can meet the trustor's expectations in the domain of decision-making. However, when algorithms are opaque, it is difficult to satisfy the expectations because people expect algorithmic decision-making to be fair, unbiased, and non-discriminatory. To fulfill the trustor's expectation, the fairness of the decision-making processes must be ensured, necessitating explicability. How, then, can a socio-technical system, which incorporates ML algorithms as a component, meet the responsiveness requirement? It is not only essential for the human agents involved to be trustworthy, but the evaluative system for assessing the ML algorithms should be validated. The algorithms should be explicable to ensure that the evaluative system is responsive to the trustor's values and interests.

If the algorithm is opaque, it is impossible to discern what bases the decision made. Consequently, we lack an understanding of why it generates its outcomes and whether they are acceptable and justifiable. Therefore, the decisions generated from the opaque processes do not guarantee trustworthiness. To build trustworthy AI, it is required to make the ML algorithms explicable because explicability is a crucial means to examine the fair and justifiable application of AI.[16]

[16] While we have been focusing on the warranted trust of decision-recipients in algorithms, the trust of algorithm users (decision-makers) matters too. They are those who use the algorithm in morally significant cases. If they do not have warranted trust in it, they are not likely to use it. Furthermore, as they make an impact on the recipients of the decision, they must be held accountable to the recipients. Thus, decision-makers also care about the trustworthiness of ML algorithms.

(3) Undermined Autonomy Calls for Explicability

Opacity undermines autonomy. The principle of autonomy in AI ethics concerns a balance between human-led and machine-led decision-making (The Montreal Declaration for Responsible AI, 2017) or "between the decision-making power we retain for ourselves and that which we delegate to artificial agents" (Floridi and Cowls, 2019). Humans have to decide "whether to delegate decisions to AI systems, to accomplish human chosen objectives" (Asilomar AI Principles, 2017). According to Floridi and others, the principle of autonomy states that "humans should retain the power to decide which decisions to take: exercising the freedom to choose where necessary and ceding it in cases where overriding reasons, such as efficacy, may outweigh the loss of control over decision-making" (Floridi and Cowls, 2019).

When it comes to opaque algorithms, humans may lose control to decide. If we lack a good grasp of how the model works, it becomes very difficult to consider good reasons to choose whether to use it (i.e., whether to accept or reject the decision suggested by the system). Consequently, the autonomy of decision-makers using ML algorithms is compromised. In line with this spirit, Robbins (2019) asserts that the principle of explicability is primarily for maintaining meaningful human control over ML algorithms. Here, the expression "meaningful human control" originates from the discussion around autonomous weapon systems and refers to the control human operators have over algorithmic uses of weapons. By generalizing this notion, Robbins (2019) adopts a specific conception of meaningful human control as "giving humans the ability to accept, disregard, challenge or overrule an AI algorithm's decision" (p. 496).

Suppose that an ML algorithm for medical diagnosis predicts that the symptom indicates skin cancer while giving no explanation. Further, suppose that your doctor does not understand the functioning of the model and when to rely on it (or when not to). Even if the algorithm is more accurate in diagnosing diseases than humans, the doctor might lack the control to decide. Doctors, as algorithm-users, have responsibility for their use to the patients. If the algorithm is not explicable to the doctor who accepts the model's decision that it is cancer, how can she be held responsible for her patients? As algorithm users make a decision and hold accountable for the decision-recipients who are affected, they need to know how the outputs are generated.

While Robbins correctly points out that explicability is for meaningful human control over ML algorithms, his assertion that autonomy is the sole objective of making ML algorithms explicable is misguided. It is true that explicability is instrumental to achieving autonomy, but this is only part of the whole story. As previously demonstrated, trustworthiness is another crucial goal of the explicability principle. Trustworthiness and autonomy are not mutually exclusive goals, but they are complementary ones. In summary, the principle of explicability has dual objectives: trustworthiness and autonomy.

5. The Dual Nature of Explicability

The remaining question is to whom the explication is owed. Given the dual goals of explicability, I contend that explication should be directed toward both decision-recipients for their warranted trust in ML algorithms and decision-makers for their informed decisions.

First, concerning the objective of trustworthiness, the whom-question can be rephrased as follows: who are the trustors to whom algorithms should be trustworthy? The decision-recipients affected by algorithmic decision-making are the trustors in question. Consequently, the recipient-oriented view on the whom-question is well aligned with the trust-based view on the goal of explicability. ML algorithms are used to assess the personal capacities or characters of decision-recipients. For a decision-making system utilizing ML algorithms to be considered trustworthy, it must fulfill the recipient's expectations and remain responsive to the decision-recipients by operating on behalf of their interests and values. When ML algorithms are used in morally significant situations, it is essential to guarantee that the decision-recipients' trust in the algorithms is warranted. Therefore, there is a good sense in which explication should be given to the decision-recipients.

Second, when it comes to retaining autonomous decisions, ML algorithms should be explicable to the decision-makers employing them. In order to make informed and responsible decisions, algorithm-users should be able to decide whether to accept, modify, or reject the recommendations generated by the algorithms. To achieve this, they must have an understanding of algorithms' functionality, capabilities, and limitations. Providing the explication of how they work enables decision-makers to make more informed and autonomous decisions, particularly when their decisions have consequences of ethical significance. Thus, the autonomy-based view informs us that explication should be directed toward the decision-makers.

Before we conclude the duality of explicability, a critical examination of Robbins' criticism of the recipient-oriented view is in order. Without any justification, Robbins (2019) assumes that the explicability principle has only one goal - maintaining meaningful human control over algorithmic decisions. He posits that with an explanation of the algorithm's decision, human beings can retain their control to decide. Although he is aware of "the ethical issues of ensuring that the outputs of algorithms are not made based upon ethically problematic or irrelevant considerations" (p. 501), he regards it as an aspect of meaningful human control. Consequently, he maintains that "an explanation of the algorithm's decision can allow for someone to accept, disregard, challenge, or overrule the rejection. This gives meaningful control of the decision to human beings"(ibid.). It would make sense if the "human beings" refer to the decision-recipients. However, Robbins contends that the explanation should be given to the decision-makers instead. While he concedes that a decision-recipient "subject to the algorithm's outputs may be interested to know the explanation," he asserts that her interest "does not establish meaningful human control over the algorithm's output" (pp. 502–503).

Robbins claims that explanations are not helpful to the decision-recipients. For example, in the case of a loan application rejected by an algorithm, the explanation of rejection may include a high debt-to-income ratio of the applicant. Without relevant domain knowledge, the applicant would be unable to assess whether the debt-to-income ratio considered was justifiable ground for rejection. Thus, Robbins concludes that the explanation fails to establish meaningful human control, and as a result, explanations need not be directed toward decision-recipients. However, this line of thought is flawed. First, as we have seen, the decision-maker's control to decide is not the only purpose of explicability. Second, our claim is not that explanations given to the decision-recipients are always sufficient for them to judge whether the algorithmic decisions are based on unethical or problematic considerations. For such judgments, decision-recipients need to have relevant information and background knowledge. Nonetheless, it does not undermine our claim that explicability is necessary for decision-recipients.[17] Most of the literature on explicability has mistakenly assumed that explication should be provided to just one stakeholder role for one and only purpose. By rejecting this unjustified assumption, we can embrace the dual nature of explicability. The duality of explicability is two-folded: first, explication serves the dual purposes of trustworthiness and autonomy; second, explanations should be directed to decision-recipients as well as decision-makers. Moreover, these two dimensions seem to overlap coherently, as explanations are needed to warrant decision-recipient's trust in ML algorithms and facilitate informed decisions by decision-makers.

6. Conclusion

Although many authors have made significant contributions to answering why ML algorithms should be explicable and to whom, they have failed to construct a coherent, systematic framework for conceptualizing and implementing explicability. In this paper, I have attempted to fit the pieces of the puzzle together. First, I argued that explicability is instrumental in enhancing more fundamental values like autonomy and trustworthiness. Second, I highlighted the dual nature of explicability, particularly aligning the answers to why- and whom-questions along the two dimensions. Explicability is directed toward both decision-makers for their autonomous and informed decisions and decision-recipients for their warranted trust in the algorithms. The next question we must address is how different explicatory strategies should be employed for different stakeholders, depending on the purposes of explicability. I suspect that intelligibility, the epistemic sense of explicability, can be met by pursuing an objectual understanding of how

[17] Robbins (2019), in footnote 12, mentioned the goal of actionable recourse, "the ability to contest incorrect decisions or to understand what could be changed in order for the data subject to achieve a more desirable result" but failed to consider further implications (p. 503).

the models work (for algorithm-users) or a causal/counterfactual explanation of why a particular decision was made (for decision-recipients). However, that is a task I aim to tackle in another article.

Acknowledgements

Earlier drafts of this article were presented at the lunchtime talk of the Center for Philosophy of Science at the University of Pittsburgh, the AI Policy Forum of the Korean Association of AI and Law, and the 10th Biennial Meeting of the Asia-Pacific Philosophy of Science Association. I am especially grateful to Edouard Machery, Collin Allen, Haksoo Ko, Gunoo Kim, and the participants for their valuable comments.

Funding

This research was supported by the Global-LAMP Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. RS-2023-00301976) and Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (RS-2024-00343989, Enhancing the Ethics of Data Characteristics and Generation AI Models for Social and Ethical Learning).

References

Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58: 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Asilomar AI Principles. 2017. *Future of Life Institute*. https://futureoflife.org/ai-principles.

Montreal Declaration for a Responsible Development of Artificial Intelligence. 2017. Announced at the conclusion of the Forum on the Socially Responsible Development of AI. https://www.montrealdeclaration-responsibleai.com/the-declaration

Baier, Annette. 1986. Trust and Antitrust. *Ethics,* 96(2), 231–260.

Beauchamp, Tom and Childress, James. 2012. *Principles of Biomedical Ethics* (7th ed). New York: Oxford University Press. https://doi.org/10.1093/occmed/kqu158

Binns, Rueben. 2018. Algorithmic Accountability and Public Reason. *Philosophy & Technology*, *31*(4), 543–556. https://doi.org/10.1007/s13347-017-0263-5

Burrell, Jenna. 2016. How the Machine' Thinks:' Understanding Opacity in Machine Learning Algorithms. *Big Data & Society.* https://doi.org/10.1177/2053951715622512

Carnap, Rudolph. 1950. *Logical Foundations of Probability.* Chicago: University of Chicago Press.

Carnap, Rudolph. 1955. Meaning and Synonymy in Natural Languages. *Philosophical Studies*, 7, 33–47.

Creel, Kathleen. 2020. Transparency in Complex Computational Systems. *Philosophy of Science*, *87*(4), 568–589. https://doi.org/10.1086/709729

European Commission. 2019. *Ethics Guidelines for Trustworthy AI.* [online] Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, et al. 2018. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* 28: 689–707. https://doi.org/10.1007/s11023-018-9482-5.

Floridi, Luciano, and Josh Cowls. 2019. A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review.* https://doi.org/10.1162/99608f92.8cd550d1

Grote, Thomas, and Philipp Berens (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics*, *46*(3), 205. https://doi.org/10.1136/medethics-2019-105586

Goodman, Bryce, and Seth Flaxman. 2017. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation." *AI Magazine*, *38*(3), 50-57. https://doi.org/10.1609/aimag.v38i3.2741

Jones, Karen. 2012. Trustworthiness. *Ethics*, 123(1), 61–85. https://doi.org/10.1086/667838

Kim, Tae Wan, and Bryan R. Routledge. 2021. Why a Right to an Explanation of Algorithmic Decision-Making Should Exist: A Trust-Based Approach. *Business Ethics Quarterly*, 1–28. https://doi.org/10.1017/beq.2021.3

Langer, Markus, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296: 103473. https://doi.org/10.1016/j.artint.2021.103473.

McLeod, Carolyn. 2021. Trust. *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2021/entries/trust/>.

London, Alex John. 2019. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report*, *49*(1), 15–21. https://doi.org/10.1002/hast.973

Lu, Joy, Dokyun (DK) Lee, Tae Wan Kim, and David Danks. 2019. Good Explanation for Algorithmic Transparency. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3503603.

Nadella, Satya. 2016. Microsoft's CEO explores how humans and AI Can solve society's challenges— together. Slate. https://slate.com/technology/2016/06/microsoft-ceo-satya-nadella-humans-and-a-i-can-work-together-to-solve-societys-challenges.html

Nickel, Philip J., Maarten Franssen, and Peter Kroes. 2010. Can We Make Sense of the Notion of Trustworthy Technology? *Knowledge, Technology & Policy* 23: 429–444. https://doi.org/10.1007/s12130-010-9124-6.

OECD (Organisation for Economic Co-operation and Development). 2019. *Recommendation of the Council on Artificial Intelligence*. Available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York: Penguin Random House.

Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*, Cambridge, MA: Harvard University Press.

Rieder, Gernot, Judith Simon, and Pak-Hang Wong. 2021. Mapping the Stony Road toward Trustworthy AI: Expectations, Problems, Conundrums. In M. Pelillo and T. Scantamburlo (eds.) *Machines We Trust: Perspectives on Dependable AI*, The MIT Press. https://doi.org/10.7551/mitpress/12186.003.0007

Ribeiro, Marco Tulio, Sammer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, 1135–1144. https://doi.org/10.1145/2939672.2939778

Robbins, Scott. 2019. A Misdirected Principle with a Catch: Explicability for AI. *Minds and Machines*, *29*(4), 495–514. https://doi.org/10.1007/s11023-019-09509-3

Roose, Kevin. 2023. How Does ChatGPT Really Work? *New York Times* (March 28, 2023) https://www.nytimes.com/2023/03/28/technology/ai-chatbots-chatgpt-bing-bard-llm.html

Simon, Judith. 2013. Trust. In: D. Pritchard (Ed.), *Oxford Bibliographies in Philosophy*. New York: Oxford University Press. Available online at: https://www.oxfordbibliographies.com/view/document/obo-9780195396577/obo-9780195396577- 0157.xml

Organon F 32 (4) 2025: 402–420 ISSN 2585-7150 (online)
https://doi.org/10.31577/orgf.2025.32402

RESEARCH ARTICLE

Debates on Semantic Relationism: Against the Psychological Variant

Seong Soo Park[*]

Received: 2 December 2024 / Revised: 25 May 2025 / Accepted: 30 May 2025

Abstract: This paper deals with the debates on whether semantic relationism can adequately address belief reports. Soames argues that semantic relationism encounters difficulties in addressing belief reports, to which Fine provides a reply. However, Pinillos contends that Fine's responses are problematic and proposes an alternative theory that, he claims, can handle belief reports without issues. The primary aim of this paper is to argue against Pinillos's claims. I begin by discussing the Fine-Soames debate on semantic relationism, focusing on the propositions suggested by Fine. I then introduce Pinillos's criticism of Fine's response to Soames's objection and argue that this criticism is inconclusive. Following that, I outline Pinillos's alternative view and demonstrate that it faces a serious problem. Finally, I conclude by summarizing the main points discussed throughout the paper.

Keywords: Semantic Relationism; belief ascriptions; Frege's puzzles; Kit Fine; Ángel Pinillos.

[*] Sungkyunkwan University
https://orcid.org/0000-0002-9626-5817
✎ The Department of Philosophy, Sungkyunkwan University, 25-2, Sungkyunkwan-ro, Jongro-gu, Seoul, Republic of Korea
✉ ysys1877@skku.edu

1. Introduction

According to semantic relationism, there is a semantic relation that holds between expressions.[1] Semantic relationists postulate such relations to explain the cognitive difference between identity claims containing co-referential terms. For example, they argue that the cognitive difference between the sentences "Hesperus is Hesperus" and "Hesperus is Phosphorus" can be explained by the fact that their semantic relations differ. While the two occurrences of "Hesperus" in the former sentence are positively (or strongly) coordinated with each other, the occurrence of "Hesperus" and the occurrence of "Phosphorus" in the latter are negatively (or weakly) coordinated.[2] On their view, the information that the two occurrences of "Hesperus" co-refer to the same object is semantically encoded in the former sentence, but the one that "Hesperus" and "Phosphorus" co-refer is not so in the latter; in the relation between "Hesperus" and "Phosphorus," the fact that they co-refer is merely empirically, not semantically, confirmed.[3]

[1] See, e.g., Putnam (1953), Tascheck (1995; 1998), Fine (2007; 2010), Pinillos (2011; 2015), Gray (2017; 2022), and Yoon (2021; 2022).

[2] The term "coordination relation," sometimes also referred to as "de jure coreference," does not necessarily refer to a *semantic* relation. Indeed, different philosophers define this term in various ways. For different interpretations of this term, see Salmon (2012), Schroeter (2012), Heck (2014), Recanati (2012; 2016), Contim (2016), Bonardi (2019), and Lee (2019).

[3] The anaphoric relation between a proper name and a pronoun can serve as a good example of a positive coordination relation, just as the relation that typically holds between repeated occurrences of the same name in ordinary discourse. Consider the sentence: "Sally sent an email to Tom, but he didn't reply." In this sentence, "Tom" and "he" are typically interpreted as standing in a positive coordination relation; a competent speaker who understands it must recognize that the two expressions co-refer. In contrast, just as two distinct yet co-referential names can illustrate a negative coordination relation, so too can a name and a co-referential pronoun used demonstratively. Consider the sentence: "When Cicero was consul, he [pointing at someone who happens to be Cicero] denounced Catiline" (Goodsell 2014, 292). Although "Cicero" and "he" in this sentence co-refer to the same individual, they are negatively coordinated. Because the pronoun is used demonstratively, someone who understands the sentence may wonder whether "Cicero" and "he" in this sentence refer to the same individual.

Semantic relationists have two logically possible options regarding *where* coordination relation obtains. The first option is to postulate that coordination holds in a piece of concrete (or usual) discourse which is characterized by the continuity of time and place in which sentences are uttered. The second option is to postulate that coordination holds in the universal discourse where all utterances and inscriptions ever made are included. If semantic relationists adopt the first option, then coordination does not hold across different pieces of concrete discourse. In contrast, the second option implies that coordination always holds in the universal discourse.

Kit Fine (2010) adopts the second option to avoid Scott Soames's objection (2010).[4] Soames argues that semantic relationism cannot deal with belief reports involving three different co-referential terms. Fine partially concedes this objection but claims that his view can resolve the raised issue by adopting the notion of a token proposition based on the second option. However, despite Fine's efforts, Ángel Pinillos (2015) claims (i) that this notion generates a new problem, and (ii) that, without relying on it, his alternative view can address belief reports in a way that avoids the problem raised by Soames.

In this paper, my goal is to move the debate on semantic relationism forward by challenging the two claims made by Pinillos. While his view is a relatively recent addition to the debate on semantic relationism, it has so far been left largely undiscussed. So, I believe that a close examination of his view, which could provide valuable insights, is worth pursuing.

The paper is structured as follows. In Sections 2 and 3, I discuss the Fine-Soames debate on semantic relationism, with a focus on the propositions suggested by Fine. Then, in Section 4, I introduce Pinillos's criticism of Fine's response to Soames's objection and argue that Pinillos overlooks the possibility that a subject can assent to a sentence without exactly grasping the content of the sentence to which they are assenting. In Sections 5 and 6, I outline Pinillos's alternative view and argue that it encounters a serious problem. Finally, in Section 7, I conclude by summarizing the main points discussed throughout the paper.

[4] It is rather unclear whether Fine (2007) adopts the second option from the outset in his initial work.

2. Coordinated Propositions and Soames's Objection

According to the basic idea of semantic relationism, a coordination relation should be reflected in the semantic content of a sentence because the relation is semantic. To articulate this idea, Fine (2007) contends that the semantic content of a sentence is a coordinated proposition. A coordinated proposition, according to Fine, can be obtained by adding a coordination relation to a singular proposition, where the semantic content of a name in a singular proposition is simply its referent. Consider the following two sentences:

(1) Hesperus is Hesperus.
(2) Hesperus is Phosphorus.

While (1) and (2) express the same singular proposition that can be represented by <Venus, Venus, being identical>, they express different coordinated propositions. This is because the names in (1) and (2) have different coordination relations; as mentioned in Sect. 1, the occurrences of "Hesperus" in (1) have a positive (strong) coordination relation, whereas the occurrences of "Hesperus" and "Phosphorus" in (2) have a negative (weak) coordination relation.[5]

However, for (1) and (2) to express different contents in this way, a coordination relation must also hold between *individuals* in singular propositions in a way that corresponds to the coordination holding between expressions.[6] Consequently, the coordinated propositions expressed by (1) and (2) can be represented by (1p) and (2p), where the bold and wavy lines represent a positive and negative coordination relation, respectively:

[5] Some philosophers might think that having a negative coordination relation means that there is no coordination relation between two expressions. I disagree with this understanding. This is because such an understanding entails that knowing a coordinated proposition with a positive coordination guarantees the knowledge of the corresponding proposition with a negative coordination. For example, on such an understanding, knowing the content of (1) implies knowing the content of (2). One way to block this undesirable consequence is to understand that a negative coordination relation is itself a relation, such as a weak coordination.

[6] On this point, Fine (2007, 54) says, "Differences in semantic relationship between names will actually show up as differences in content."

(1p) <$\underline{\text{Venus}}$, $\underline{\text{Venus}}$, being identical>

(2p) <$\underset{\sim\sim\sim}{\text{Venus}}$, $\underset{\sim\sim\sim}{\text{Venus}}$, being identical>

One characteristic that might be considered an advantage of adopting coordinated propositions is that it seems to enable semantic relationists to address the presumed semantic difference between belief reports involving co-referential names. Consider the following sentences:

(3) Tom believes that Hesperus is Hesperus.

(4) Tom believes that Hesperus is Phosphorus.

Intuitively, sentences (3) and (4) can be *judged* as semantically different by ordinary people, especially when Tom agrees with (1) but disagrees with (2). Semantic relationists can explain why ordinary people make such judgments if they treat coordinated propositions as the objects of our beliefs.[7] According to this strategy, the reason why ordinary people may judge that (3) and (4) are semantically different is that the sentences are indeed semantically different: the coordinated proposition expressed by the that-clause in (3) and the one expressed by the that-clause in (4) are different as shown in (1p) and (2p). Thus, at first glance, semantic relationists can deal with belief reports involving co-referential names in a way that respects ordinary people's intuition.

Despite this initial attraction, assuming coordination occurs in a piece of concrete discourse, Soames argues that coordinated propositions understood as the objects of our belief pose a problem for semantic relationists. According to him, the problem becomes evident in belief reports involving three co-referential terms.[8] To illustrate this, consider the following sentence:

[7] Fine (2007) denies that coordinated propositions are the objects of our beliefs. In his initial work, given the Paderewski case, he argues that it is nearly impossible to determine which type of proposition can be regarded as the object of belief. Thus, this articulation merely illustrates one possible approach a semantic relationist might take in dealing with belief reports. However, in his later work (2010), Fine admits that he at least needs to propose a proposition corresponding to the objects of belief.

[8] As mentioned in footnote 5, this objection is somewhat misleading, as Fine (2007) clearly denies that coordinated propositions are the objects of our beliefs. Nevertheless, it motivates Fine (2010) to address the question, "Then, which type of proposition is the

(5) Venus is Phosphorus.

Now, suppose that (2) and (5) are uttered in different pieces of concrete discourse. In this case, the occurrence of “Hesperus” and the occurrence of “Phosphorus” in (2) are negatively coordinated with each other. But the same applies to (5). Even a competent language user may ask whether the occurrence of “Venus” and the occurrence of “Phosphorus” in (5) refer to the same object. This suggests that their relation does not semantically indicate that they refer to the same object. As a result, (2) and (5) express the same coordinated proposition. The problem is that this forces semantic relationists to accept that the following belief report is true when belief report (4) is true:

(6) Tom believes that Venus is Phosphorus.

This consequence is undesirable for semantic relationists because they already admit that (3) and (4) are semantically different. Yet, they cannot maintain this distinction for (4) and (6) because the coordinated proposition expressed by the that-clause in (4) is the same as the one expressed by the that-clause in (6). Thus, when (4) is true, (6) has to be true. However, ordinary people may still judge that (4) and (6) are semantically different, especially when Tom agrees with (2) while disagreeing with (5).

3. Fine’s Reply: Introducing Token Propositions

To avoid the unwanted result that (4) and (6) are semantically the same, Fine (2010) clarifies his view by stating that he adopts the second option—that coordination occurs in the universal discourse rather than in a piece of concrete discourse. This implies that Fine must propose a more complex picture of propositions than the one shown in (1p) and (2p), where coordination seems to be assumed as holding in a local piece of discourse. Based on this motivation, Fine (2010) introduces a new type of proposition, namely, a token proposition.

object of our belief?” Consequently, Fine acknowledges a theoretical gap in his theory of belief reports. We will examine Fine’s response to this objection in the next section.

Token propositions are structured entities that consist of token individuals, token properties, and token relations. According to Fine (2010), there is a universal body, say D, that contains all *occurrences* of singular propositions ever expressed. Fine contends that the numerical identity of a token individual, token property, and token relation is determined by the equivalence classes of positive coordination in D.

To illustrate this, imagine that the only uttered sentences throughout human history are (1), (2), and (5), and they have been uttered many times. In this case, there are three numerically different token individuals. This is because there are three equivalence classes of positive coordination in D: one corresponds to occurrences resulting from the uses of "Hesperus," another to those resulting from "Phosphorus," and the last to those resulting from "Venus." These token individuals are different from ordinary individuals and are thus commonly viewed as abstract objects. Thus, I will represent token individuals in italics to distinguish them from ordinary individuals. As a result, we can say that there are three token individuals, namely, *Hesperus*, *Phosphorus*, and *Venus*.

Although the whole picture may seem complex, the underlying idea is simple: if there are N (where 'N' to be replaced with a number) legitimate lexical items that have referents, there are N numerically different token individuals as long as there are no synonyms.

In what follows, the numerical identity of a token individual, token property, and token relation determines the numerical identity of a token proposition. For example, suppose that John utters, "Hesperus is a planet" following his utterance of "Phosphorus is a planet." In this case, the token propositions expressed by John's two utterances are numerically different because they contain numerically different token individuals, *Hesperus* and *Phosphorus*. But, if John utters "Hesperus is a planet" again after uttering the previous two sentences, then the token proposition expressed by this utterance is numerically the same as the one expressed by John's earlier utterance of "Hesperus is a planet."

It is worth noting that two numerically different token propositions are semantically different. This is because their numerical distinctiveness implies that one of their constituents stands in different semantic relations with others. As a result, by adopting token propositions, Fine can address Soames's objection: assuming that token propositions may be construed as the objects of our beliefs, (6) does not have to be true when (4) is true. This is because the token propositions expressed by (4) and (6) are numerically different. The one expressed by the that-clause in (4) contains *Hesperus* and *Phosphorus*, whereas the one expressed by the that-clause in (6) contains *Venus* and *Phosphorus*. Therefore, on this view, sentences (4) and (6) may differ in their truth value, which aligns with ordinary people's judgement.

4. Pinillos's Criticism of Token Propositions

Pinillos (2015) argues that although the strategy of adopting token propositions helps semantic relationists avoid Soames's objection, it generates a new problem. To argue for this, Pinillos (2015, 332) asks us to consider the following situation, assuming that token propositions are what belief subjects believe as Fine does:

Peter utters "Hesperus is a planet" at time t_1. After Betty overhears Peter's utterance, she accepts it right away because she fully trusts him. So, Betty self-reports it sincerely by saying, "I believe that Hesperus is a planet." After a while, Peter utters "Hesperus is a planet" once again at t_2. Even though Betty has no idea whether the two occurrences of the name "Hesperus" in Peter's utterances refer to the same object, she also accepts his second utterance after overhearing it, and sincerely makes another self-report by saying, "I believe that Hesperus is a planet."

Let's call Peter's first and second utterances *A* and *B*, and Betty's two self-belief reports *P* and *Q*, respectively. Pinillos argues that *A*, *B*, and the that-clauses in *P* and *Q* express numerically the same token proposition. This is despite the fact that Betty does not know whether the two instances of the name "Hesperus" in Peter's utterances refer to the same object. This is because, according to Pinillos, Betty intends to refer to the same object as the one Peter wants to refer to, and all of Peter's uses of "Hesperus" indeed refer to the same object, Venus. Thus, Pinillos concludes that *A*, *B*, and the that-clauses in *P* and *Q* express not only the same singular proposition but also numerically the same token proposition.

Now, Pinillos argues that this consequence raises a problem. The reason is that, according to his description, Betty herself thinks that her belief reports P and Q are about her two distinct beliefs. Pinillos takes this to conflict with Fine's view. That is, from Pinillos's perspective, if A, B, and the that-clauses in P and Q all express numerically the same token proposition, then it cannot be the case that Betty believes that her self-reports are about two distinct beliefs.

It is worth noting that, for this line of objection to work, it must be the case that Betty believes the token propositions expressed by Peter's two utterances (i.e., A and B) when she sincerely assents to them. Indeed, since the token propositions expressed by A and B are numerically the same, if she believes the token propositions expressed by the sentences to which she sincerely assents, then she cannot believe that P and Q are about two distinct beliefs, so long as she is rational. The principle that licenses this move is the disquotational principle, which connects sincere assent with belief: if a normal speaker, on reflection, sincerely assents to the sentence "S," then she believes that S (Kripke 1979). Since Betty sincerely assents to A and B, *if* the disquotational principle applies in this case, it seems to follow from Fine's view that she cannot believe that her belief reports are about her two distinct beliefs, which is contrary to what actually happens.[9]

[9] Pinillos also appears to recognize that his objection significantly weakens if Fine rejects the disquotational principle. One potential problem is that, as the reviewer notes, Fine explicitly states that he does not aim to offer a fully compositional semantics—though he may still aim to do so to a reasonable extent without fully abandoning the principle. Despite this, Pinillos maintains that his objection "at the very least" poses a problem even in this case, since the two occurrences of "Hesperus" in P and Q should be negatively coordinated, given that Betty could wonder whether they refer to the same object, yet P and Q express the same token proposition on Fine's view. However, I do not think this threatens Fine's view. First, even if Betty could raise such a question, it would make no sense for Peter to do so, and Betty defers to Peter's referential intentions in her reports. Second, Fine now holds that coordination obtains in universal discourse, thereby abandoning the view that coordination is non-transitive. While this conflicts with his earlier position (2007), it does not undermine his current proposal (2010) unless further reasons are given. Thus, I will consider whether Pinillos's objection can succeed under the assumption that Fine could endorse the disquotational principle, except for hard cases like the Paderewski case. (Fine does in fact seem to do so in developing his notion of token proposition, at least when responding to Soames's objection.)

However, I do not think that Pinillos's criticism refutes the idea that token propositions are the objects of our beliefs. This is because the disquotational principle does not hold in the Betty and Peter case. More specifically, the principle applies only to cases where a subject clearly understands what she sincerely assents to. And this seems like a reasonable constraint: a subject cannot believe what she does not understand. In fact, a recent version of the disquotational principle already adopts this constraint explicitly: if S (a sentence) means P (a proposition) and a speaker *understands* and sincerely accepts S, then the speaker believes P (Speaks 2010).

The important point is that, in this case, Betty does not understand the content expressed by the sentences to which she assents. This is evident for two reasons. First, as mentioned, Peter's two utterances express numerically the same token proposition. So, if Betty really understood the content of the sentences to which she assented, she would have recognized that her belief reports are about the same belief. However, according to Pinillos's description, this is not the case. Thus, Betty's judgment about her belief reports provides good reason to think that she does not understand the content of the sentences to which she assents. Second, and more straightforwardly, Betty is partially ignorant of the content of the token proposition expressed by Peter's two utterances: she has no idea whether the referent of "Hesperus" in A is the same as the referent of "Hesperus" in B.

Now, if Betty does not understand what she assents to, then there is no reason to accept that it must be the case that Betty believes the token propositions expressed by A and B merely because she sincerely assents to them. This is because the disquotational principle cannot be applied in this case. But if so, then there is no problem at all with Betty believing that P and Q are about her two distinct beliefs. The lesson from this is that it is one thing to say that a sentence or belief report uttered by someone expresses some semantic content, but it is another thing to say whether she stands in a position to know what such semantic content is.

Before ending this section, let me briefly compare the Betty and Peter case with the Paderewski case in order to clarify the structure of Pinillos's objection and its limitations. The two cases are similar in that they are both intended to generate a difficulty if the disquotational principle holds. According to one well-known version of the Paderewski case, a subject assents to the sentence “Paderewski had musical talent” and also to the sentence “Paderewski had no musical talent,” while being unaware that the name “Paderewski” in both sentences refers to the same person. Since the subject is assumed to be rational, if the disquotational principle holds in this case, it leads to a problematic result—namely, that a rational person can hold contradictory beliefs.

However, the difference is that, in the Paderewski case, it is assumed that the subject understands the contents expressed by the sentences to which they assent. This is because the propositions the subject is supposed to grasp in the case are singular propositions, not token propositions, and thus we can say that the subject does in fact understand them, since the subject is acquainted with Paderewski when they assent to the two sentences. The subject simply does not know that Paderewski under one mode of presentation is the same person as Paderewski under another. Thus, the Paderewski case genuinely challenges the notion of singular proposition, unless one gives up the disquotational principle.

The Betty and Peter case has exactly the same structure: it is originally intended to challenge the notion of token proposition, unless one gives up the disquotational principle. However, the problem is that, in the Betty and Peter case, Betty does not understand the content expressed by the sentences to which she assents, as she does not even know what the referent of the two occurrences of “Hesperus” is. Therefore, the disquotational principle cannot be applied in this case, and thus the case does not pose a problem for Fine’s notion of token proposition.

5. Pinillos’s Alternative View

After presenting his criticism of the adoption of token propositions, Pinillos suggests an alternative view on belief reports that a relationist might adopt. He believes that without introducing the universal body or discourse, presumably retaining the first option that coordination holds in a concrete piece of discourse, coordinated propositions alone can address Soames’s objection to semantic relationism. In this section, I will outline Pinillos’s alternative view before arguing that it still faces several problems.

The key idea of Pinillos's view is that when we evaluate a belief report, we should consider what is presupposed by the speaker in the discourse where the belief report is uttered and incorporate it into a semantic content. According to Pinillos, there are two types of presupposition that the speaker of a belief report might have: one is specific, the other is non-specific. Under the specific presupposition, the ascriber (i.e., a speaker) presupposes that the ascribee (i.e., the target agent of a belief report) is thinking of an object in a specific way that relates to a definite description. Under the non-specific presupposition, the speaker simply presupposes that the ascribee is thinking of an object in some way. Pinillos argues that considering these presuppositions helps a relationist view address problems related to belief reports.

Taking Soames's idea (2002) as motivation, Pinillos (2015, 333) argues that if a speaker involves a specific presupposition, then a belief report made by the speaker not only expresses a semantic content, but also conveys a content resulting from what is asserted by it. Following Soames, Pinillos calls the latter type of content a *descriptively enriched proposition.* According to Pinillos, descriptively enriched propositions may reflect the speaker's specific presupposition and, thus, may contain the mode of presentation of an object as part of its content.

For example, suppose that John presupposes that Mary is thinking of Venus as the brightest object visible in the morning sky. Also, suppose that Mary does *not* know Hesperus is Phosphorus. Now, John utters the following sentence:

(7) Mary believes that Phosphorus shines.

According to Pinillos's view, the contents associated with a sincere *de dicto* use of (7) can be represented by (7p) plus (7pr), where (7p) represents what is expressed by (7), and (7pr) represents what is asserted by (7). The novel idea of Pinillos is that the occurrence of Venus in (7p) is positively coordinated with the one in (7pr), and therefore, allows (7pr) to make a *semantic* contribution to the truth condition of (7) as shown below.

(7p) <Believes, Mary, <Venus, shines>>.

(7pr) <Believes, Mary, < [the *x*: (*x* is the brightest visible object in the morning sky and *x*=Venus)], shines>>.

More specifically, Pinillos (2015, 333) claims that if this is the case, belief report (7) will be true if and only if Mary has the following set of coordinated beliefs: {<Venus, shines>, < [the x: (x is the brightest visible object in the *morning* sky and x=Venus)], shines>}. But Mary does not have this coordinated set of beliefs because she does not know that Hesperus is Phosphorus. Presumably, Mary believes that the object in question is the brightest visible object in the night sky instead of the morning sky. Thus, on this view, belief report (7) is false.

To be sure, there could be cases in which one has no idea about the specific way that the ascribee is thinking of an object. Pinillos claims that in these cases, the ascriber has a non-specific presupposition. More specifically, according to Pinillos (2015: 335), in such a situation, the ascriber may still assume that the ascribee is thinking of an object in some way with a conception of the object. For instance, let us suppose that (7) is uttered by Carol, and that Carol has no idea about the specific way that Mary is thinking of Venus. In this case, there would be no reason to believe that a definite description, such as the brightest visible object in the morning sky, is involved in Carol's presupposition. However, according to Pinillos, even in this situation, it seems reasonable to say that Carol is presupposing that Mary is thinking of Venus in some way related to a conception of Phosphorus: otherwise, Carol would not have used the term "Phosphorus" in uttering (7). Pinillos (2015, 336) claims that if so, belief report (7) involves the following truth condition: Mary thinks that Venus shines, where Mary is thinking of Venus under the conception of Phosphorus. However, given that Mary does not know that Hesperus is Phosphorus, it is not the case that she is thinking of Venus under the conception of Phosphorus. Thus, in this case, belief report (7) is also false.

It is worth noting that Pinillos's view can address Soames's objection. This is because regardless of whether a speaker uttering (6) has a specific presupposition or not, there could be cases where the truth conditions of (4) and (6) differ, especially when Tom does not know that Hesperus is Venus. Under a case involving a specific presupposition, for (6) to be true, Tom must think of Venus in a way that associates it with a definite description of "Phosphorus," such as the brightest object in the morning sky. However, the truth condition of (4) may not require this. Similarly, under a case involving a non-specific presupposition, the truth condition of (6) requires Tom to think of Venus under a conception of Phosphorus, whereas the truth condition of (4) does not. Thus, on Pinillos's view, belief reports (4) and (6) may be semantically different, which explains why ordinary people may judge them to be semantically different.

6. Problems with Pinillos's View

Although Pinillos's view might initially seem to respect ordinary people's judgement regarding belief reports, it faces several problems. One major problem is that it imposes overly strict truth conditions on belief reports, particularly when an ascriber has a specific presupposition about how the ascribee thinks of an object. To make this point clear, let me first compare Pinillos's view with Soames's.

As mentioned, Pinillos's view is partially motivated by Soames's, so their approaches to belief reports may seem similar. Indeed, Soames agrees that a sincere *de dicto* use of (7) may result in conveying a descriptively enriched proposition. However, there is a fundamental difference between their views.

The difference lies in the semantic contribution of a descriptively enriched proposition. On Soames's view, for example, what is asserted by (7) (i.e., (7pr)) does not affect the truth condition of (7). The asserted content remains at a pragmatic level. More specifically, Soames (2002) holds that belief report (7) is *true* even if Mary has not heard of the term "Phosphorus." According to Soames's view, Mary in fact believes that Phosphorus shines in one guise but is cognitively inaccessible to this fact because her belief was formed in another guise. Soames argues that the reason ordinary people judge (7) to be false is that they fail to distinguish between what is semantically expressed and what is pragmatically conveyed by a sentence. In contrast, according to Pinillos's view, what is asserted by (7), based on the relation of coordination, plays an important role in making (7) false. Pinillos (2015, 335) claims that this difference can be seen as an advantage of his approach over Soames's.

However, allowing the asserted content expressed by a belief report to contribute to its semantic value causes a similar problem to the one that arises in non-specific presupposition cases. The problem is that this makes the truth condition of a belief report too strict. This is because, according to Pinillos's view, in a case involving a specific presupposition, a belief report is true if and only if the belief report *precisely* describes ascribee's way of thinking about the object that is being considered. More specifically, consider the following situation.

Fred and Sally are taking the same course about Greek history. In the course, Fred came to have a false belief about Aristotle because he fell asleep. As a result, he ended up mistakenly associating the name "Aristotle" with the definite description "the most famous teacher of Plato." Unlike Fred, Sally correctly learned that Aristotle is the most famous pupil of Plato's. One day, Sally utters that Aristotle is a Greek philosopher. After hearing this statement, Fred makes the following belief report:

(8) Sally believes that Aristotle is a Greek philosopher.

Belief report (8) seems intuitively true. The problem with Pinillos's view is that Fred may presuppose that Sally thinks of Aristotle in a specific way. Since Fred mistakenly believes that Aristotle is the most famous teacher of Plato, and Sally is taking the same course as he does, Fred might presuppose that Sally also thinks of Aristotle as the most famous teacher of Plato. And this prevents (8) from being true. Indeed, on Pinillos's view, a sincere *de dicto* use of (8) expresses (8p) and conveys (8pr) as follows:

(8p) <Believes, Sally, <Aristotle, being a Greek philosopher>>.
(8pr) <Believes, Sally, < [the x: (x is the most famous *teacher* of Plato's and x=Aristotle)], being a Greek philosopher>>.

So, (8) will be true if and only if Sally has the following set of coordinated beliefs: {<Aristotle, being a Greek philosopher>, < [the x: (x is the most famous teacher of Plato and x=Aristotle)], being a Greek philosopher>}. However, Sally does not have this coordinated set of beliefs since she correctly learned that Aristotle is the most famous pupil of Plato. Thus, Pinillos's view implies that (8) is false, which is counterintuitive.[10]

[10] Since Pinillos's view is partially based on Soames's, one might suspect that the same problem arises in Soames's view. This suspicion may stem from the fact that, on Soames's view, (8) is semantically true but may count as false at the assertive

Regarding this problem, one might argue that even though this kind of truth condition is strict, it makes perfect sense when a belief report is involved in a sincere *de dicto* use*,* not *de re* use. However, I disagree with this. Note that I have never argued that definite descriptions cannot be counted as a semantic factor, nor that only referents are semantic contents for names. Rather, my point is that the suggested truth condition for (8) requires us to investigate one's psychological states, which she personally associates with a singular term. And this is highly undesirable, given that the meaning of an expression is supposed to be public. Indeed, a semantic content should be communicative so that the hearer can easily grasp it. However, Pinillos's view conflicts with this idea: if the truth condition of a sentence depends on one's presupposition, this may make its content entirely inaccessible to the hearer.

Another major problem with Pinillos's view is that the truth condition of a belief report depends on what kind of presupposition the ascriber makes. More specifically, Pinillos's view requires that the truth condition of (8) depends on how specific Fred's presuppositions about Sally are. If Fred presupposes that Sally associates Aristotle with a particular definite description, then the truth condition of (8) includes the content of that definite description as a core element. On the other hand, if Fred believes that he is ignorant of the specific way Sally thinks of Aristotle, then the truth condition depends not on the content of a definite description but on

level. However, I do not think this is problematic, because what is asserted in (8), according to Soames, merely reflects what the speaker intends to convey; the fact that (8) is false at the assertive level, on Soames's account, simply shows that what the speaker presupposes is false.

Here, it is worth noting that this does not imply that ordinary people will necessarily take (8) to be false. This is because they may not grasp what the speaker intends to assert in uttering the sentence. Of course, there are many cases where ordinary people clearly grasp what is asserted or even confuse it with the semantic content, as Soames notes in his account of Frege's Puzzle. But the point is that the case involving (8), with the false belief about Plato, does not belong to such cases. The problem with Pinillos's view is that it cannot allow (8) to count as false merely in terms of the speaker's presupposition. In fact, according to his view, (8) is semantically false—not merely assertively false. I am grateful to the referee for giving me the opportunity to clarify this issue.

a particular conception Fred has in mind regarding Sally's belief. But it is difficult to understand why the truth condition would have to change in this way merely based on the reporter's presuppositional specificity.

These problems point to the fundamental issue with Pinillos's view. After all, the issue lies in the fact that, on his view, it is the psychological states of the reporter, not those of the ascribee, that affect the truth condition of a belief report. Intuitively, how Fred thinks of Sally's thought about Aristotle seems irrelevant when reporting Sally's belief about Aristotle. However, Pinillos's view fails to respect this point. To put the point differently, Pinillos's view is unpromising—again, given the idea that the meaning of an expression is supposed to be public. His view clearly rejects this idea by making the truth conditions of belief reports overly sensitive to the ascriber's psychological states, thereby rendering those truth conditions ones that are inaccessible to third parties.

7. Conclusion

In this paper, I have articulated Fine's semantic relationism and the related discussions with a focus on the logical options available to Fine. I then examined whether Pinillos's argument against Fine's view and his alternative view are convincing and concluded that neither is.

More specifically, I first argued that Pinillos's criticism of Fine's view fails to convincingly show that token propositions are problematic. In my view, this is because Pinillos overlooks the crucial fact that a subject can assent to a sentence without exactly grasping its content. In addition, I argued that Pinillos's alternative view is also unconvincing, as it strongly ties the psychological states of a reporter to the semantic content of a belief report. While it is true that we express our private thoughts through sentences, this does not have to mean that any personal thought may be conflated with semantic content, given our linguistic practices in communication.

Lastly, it should be noted that this paper did not provide any arguments against Fine's view. Investigating that issue is a task for another occasion. However, I believe that my criticism of Pinillos's attempts contributes to the discussion about semantic relationism by highlighting the need for further investigation of Fine's view. In fact, if someone believes that Fine's view is unpromising, this paper may motivate them to provide a compelling argument against it and an alternative view.

Acknowledgements

I thank David Braun and Chulmin Yoon for giving me the initial motivation to write this paper. I am also grateful to an anonymous reviewer whose detailed comments helped me to improve the paper. My special thanks go to Inchul Yum, with whom I had many thoughtful discussions.

References

Bonardi, Paolo. 2019. “Manifest Validity and Beyond: An Inquiry Into the Nature of Coordination and the Identity of Guises and Propositional-Attitude States.” *Linguistics and Philosophy* 42: 475–515. https://doi.org/10.1007/s10988-018-9245-z

Contim, Filipe Drapeau V. 2016. “Mental Files and Non-Transitive de Jure Coreference.” *Review of Philosophy and Psychology*, 7: 365–88. https://doi.org/10.1007/s13164-015-0251-6

Fine, Kit. 2007. *Semantic Relationism.* Oxford: Blackwell.

Fine, Kit. 2010. “Comments on Scott Soames’ ‘Coordination Problems.’ ” *Philosophy and Phenomenological Research*, 81: 475–84. https://doi.org/10.1111/j.1933-1592.2010.00405.x

Gray, Aidan. 2017. “Relational Approaches to Frege’s Puzzle.” *Philosophy Compass*, 12: 1–15. https://doi.org/10.1111/phc3.12429

Gray, Aidan. 2022. “Minimal Fregeanism.” *Mind*, 131: 429–58. https://doi.org/10.1093/mind/fzab076

Heck, Richard. 2014. “In Defence of Formal Relationism.” *Thought: A Journal of Philosophy*, 3: 243–50. https://doi.org/10.1002/tht3.138

Kripke, Saul. 1979. “A Puzzle About Belief.” In *Meaning and Use*, edited by A. Margalit, 239–83. Dordrecht: Reidel.

Lee, Poong. 2019. “Co-filing and de Jure Co-referential Thought in the Mental Files Framework.” *Erkenntnis*, 87: 309–45. https://doi.org/10.1007/s10670-019-00196-1

Recanati, Francois. 2012. *Mental Files.* Oxford: Oxford University Press.

Recanati, Francois. 2016. *Mental Files in Flux.* Oxford: Oxford University Press.

Pinillos, Ángel. 2011. “Coreference and Meaning.” *Philosophical Studies*, 154(2): 301–24. https://doi.org/10.1007/s11098-010-9543-y

Pinillos, Ángel. 2015. “Millianism, Relationism, and Attitude Ascription.” In *On reference*, Andrea Bianchi (ed.): 322–346. Oxford: Oxford University Press.

Putnam, Hilary. 1953. “Synonymity and the Analysis of Belief Sentences.” *Analysis*, 14: 114–22. https://doi.org/10.1093/analys/14.5.114

Salmon, Nathan. 2012. “Recurrence.” *Philosophical Studies*, 159: 407–41. https://doi.org/10.1007/s11098-011-9773-7

Schroeter, Laura. 2012. “Bootstrapping Our Way to Samesaying.” *Synthese*, 189: 177–97. https://doi.org/10.1007/s11229-012-0099-6

Soames, Scott. 2002. *Beyond Rigidity: The Unfinished Semantic Agenda of Naming and Necessity*. Oxford: Oxford University Press.

Soames, Scott. 2010. “Coordination Problems.” *Philosophy and Phenomenological Research*, 81: 464–74. https://doi.org/10.1111/j.1933-1592.2010.00404.x

Speaks, Jeff. 2010. “Explaining the Disquotational Principle.” *Canadian Journal of Philosophy*, 40: 211–38. https://doi.org/10.1353/cjp.2010.0004

Tascheck, William. 1995. “Belief, Substitution, and Logical Structure.” *Nous*, 29: 71–95. https://doi.org/10.2307/2215727

Tascheck, William. 1998. “On Ascribing Beliefs: Content in Context.” *Journal of Philosophy*, 95: 323–53. https://doi.org/10.5840/jphil199895736

Yoon, Chulmin. 2021. “The Transitivity of de Jure Coreference: A Case Against Pinillos.” *Philosophical Studies*, 178: 2257–77. https://doi.org/10.1007/s11098-020-01545-5

Yoon, Chulmin. 2022. “Semantic Relationism.” In *The Routledge Handbook of Propositions*, edited by Chris Tillman, 470–89. New York: Routledge.

Organon F 32 (4) 2025: 421–456 ISSN 2585-7150 (online)
https://doi.org/10.31577/orgf.2025.32403

RESEARCH ARTICLE

An Essentialist Bimodal Interpretation of Descartes' Creation Doctrine

Andrew Tedder*

Received: 15 August 2024 / Accepted: 9 June 2025

Abstract: This paper develops and defends an essentialist bimodal (or biessentialist) interpretation of Descartes' Creation Doctrine. The two modalities express facts about essences: i-modalities express relations of compatibility/entailment as obtaining between propositions and the essences of created things, while o-modalities express such relations with God's essence. On this reading, the necessity of eternal truths should be understood as i-necessity, while the possibility with which God could have made the eternal truths false should be understood as o-possibility. I argue that this is a plausible reading of the central texts, and that it renders the creation doctrine coherent while improving on some previous accounts.

Keywords: Descartes; metaphysics; modal voluntarism; essence.

1. The Problem with the Creation Doctrine

An especially difficult problem in the interpretation of Descartes' metaphysics, and its relation to Descartes' broader philosophical work, concerns the modal status of the *eternal truths*. Descartes uses this term to refer to

* Ruhr University Bochum
https://orcid.org/0000-0002-2303-003X
Department of Philosophy I, Ruhr University Bochum, Universitätsstraße 150, 44801 Bochum, Germany
ajtedder.at@gmail.com

a class of necessary truths including, at least, truths about the *essences* of particular objects, as well as some more general necessary truths, such as truths of logic like the law of non-contradiction.[1] Essences and necessary truths play key roles in Descartes' metaphysical project in the *Meditations*, and in his scientific work more broadly, but these notions raise substantial difficulties in light of the (in)famous Cartesian thesis often called the *creation doctrine* (CD). Perhaps the simplest statement of CD is:

(CD) The eternal truths are *freely* created by God.

It is usually taken as a consequence of (CD) that God has the power to make the eternal truths false, and therefore that the eternal truths are possibly false. Kaufman (2002) provides a nice statement of the key problem with CD as a tension between the following two claims, to both of which Descartes seems committed:

(1) Eternal truths are necessarily true.
(2) Eternal truths are possibly false.[2]

It is widely accepted that a proposition is possibly false if and only if it is not necessarily true, and so (1) and (2) contradict each other. This raises the

[1] Descartes seems to take the term 'eternal truth' from Mersenne, for in the 15 April 1630 letter (AT 1:145, CSMK3, p.23), he writes first of "the mathematical truths which you call eternal," but, following the literature, I'll understand, by "eternal truths" just "necessary truths." All citations to Descartes' work are to, first, the Adams and Tannery (AT) original languages edition, and, second, to the English translations by Cottingham, Stoothoff, Murdoch, and Kenny. The first volume of the translated work (Descartes 1985) will be cited as 'CSM1,' the second (Descartes 1984) as 'CSM2,' and the third (Descartes 1991) as 'CSMK3.' Citations to original sources occurring in references to secondary literature will be altered to reflect this convention.

[2] This statement of (2) builds in the claim that CD entails that God has free control not just over which truths are necessary (eternal), but also over whether the actually eternal truths are true *simpliciter*. The latter claim is a stronger claim, and one may want, in the spirit of limited possibilism (discussed in §2), to assert, rather, a *prima facie* weaker iterative claim, namely: "it is possibly false that the eternal truths are necessarily true." One advantage of the kind of view I defend here is that it can render CD consistent even if it implies (2), in addition to this weaker claim (this is a point to which I'll return in the discussion of limited possibilism).

question: can CD be faithfully interpreted in a way which renders Descartes' total modal metaphysics consistent? Van Cleve (1994), for instance, argues that it cannot be consistently interpreted, though there is a long history of attempts. The extant attempts take a variety of tacks. Some approaches, most famously Frankfurt's (1977), seek to reject one of (1), (2), while others, including McFetridge (1990), seek to reinterpret the modal terms so that the apparent inconsistency is dissolved. My preferred approach, which is of a piece with McFetridge and, more recently, Saint-Germier (2018), invokes a distinction between multiple kinds of modality (hence the approach is *bimodal*). If the instances of 'possibly' and 'necessarily' occurring in (1) and (2) are read as expressing different kinds of modality, then these propositions may be consistent.

Bimodal accounts have been given in the literature. In fact, even Frankfurt's view involves a bimodal aspect, according to which Descartes does not claim that some *alethically* necessary propositions are *alethically* possibly false, but rather just that some *epistemically* necessary propositions are alethically possibly false (more on this in the next section). This last could happen if some propositions were possibly false, while being such that we could not conceive their possibility. While such an account gets something right about Descartes' modal metaphysics (indeed something about this story of conceivability is importantly right), a different bimodal account does a better job.

The kind of bimodal account developed by McFetridge, which I'll extend here, treats both modalities as alethic, so that while two different modal notions are at play, both concern metaphysical reality (and so neither has *just* to do with what a (human) agent can or cannot conceive). This move does seem to be a better fit for the central texts, but raises the question of how these two kinds of necessity are grounded – i.e. what features of metaphysical reality explain the difference between these two alethic modalities? There are a number of stories one could give here, but that which I defend here is *essentialist*. By claiming that Descartes is an essentialist I mean that he takes modal facts to reduce to (or be grounded in, or be true in virtue of, to be explained by…) facts about the essential properties (or 'essences') of things. Bimodalism falls out of this picture when we recognise that, for Descartes, there are two importantly different kinds of things with essences, and from there infer that the essences of these different kinds of thing can ground different kinds of necessary/possible truth. The kind of view I have in mind may be called a *biessentialist* view.

The two kinds of things I have in mind here are, first, *God* and, second, *things created by God*. The second class includes everything which is not God. I'll argue that Descartes should be understood as an essentialist, in the appropriate sense, and furthermore that this distinction between kinds of essence is textually plausible. I'll aim to provide a plausible basis for the bimodal solution to the problem of CD in Descartes' metaphysics. I won't aim to resolve all potential problems with such an account, but where there are looming difficulties, I'll suggest what seem to me promising ways to resolve these. The upshot will be a story according to which both of the modalities in the bimodal solution are grounded in objects recognised by Descartes' metaphysics.

Let me start by considering three influential extant interpretations and, in passing, reviewing some of the key passages in which Descartes discusses CD and its consequences.

2. Some Extant Interpretations

Much of the discussion surrounding CD has taken place against a background of assumptions and claims made by two influential interpreters. These are Frankfurt (1977) and Curley (1984), and their proposed interpretations are *universal possibilism* and *limited possibilism* respectively. According to the universal possibilist reading, Descartes is committed to the view that every proposition is *merely possible*, and hence that Descartes is not committed to (1) above. Alternately, the limited possibilist reading has it that while Descartes is committed to (1), but is not committed to (2) as written, but rather just to:

(2') Eternal truths are possibly possibly false.

So long as one rejects that if ϕ is possibly possible then ϕ is possible (a version of the distinctive axiom of the modal logic **S4**), (2') is not inconsistent with (1). The limited possibilist invokes *iterated modalities* in order to render CD consistent, whereas the universal possibilist invokes Descartes' (apparent) commitment to the claim that God's free creation of the world implies that nothing is necessary, and that God could bring about any state of affairs whatsoever. Almost as old as the limited possibilist line, and more important for my purposes, is the bimodal reading developed by McFetridge (1990), which seek to dissolve the apparent contradiction by reading (1) and (2) as concerning two distinct modalities.

I'll recapitulate the development of these three positions in some detail in order to bring out some of the terrain in which the discussion of CD takes place, and to trace the development of some key positions and views in the literature.

2.1. Frankfurt: Universal Possibilism

Frankfurt's interpretation is perhaps the most influential of recent accounts, though the main concern of many commentators is to avoid his conclusion: according to Descartes, every proposition is *merely* possible. To start, Frankfurt notes:

> What Descartes calls 'eternal truths' are truths about essences. The Pythagorean theorem, for example, is (or purports to be) an eternal truth about what is essential to right triangles. Asserting that the eternal truths are laid down by God is tantamount, then, to saying that God is the creator of essences. (Frankfurt 1977, p. 38).

Evidence for this essentialist reading of eternal truths is available in a letter to Mersenne, among the earliest statements of CD:

> You ask me by what kind of causality God established the eternal truths. I reply: by the same kind of causality as He created all things, that is to say, as their efficient and total cause. For it is certain that *He is the author of the essence of created things no less than of their existence; and their essence is nothing other than the eternal truths.* You ask also what necessitated God to create these truths; and I reply that He was free to make it not true that all radii of the circle are equal – just as free as He was not to create the world. And it is certain that *these truths are no more necessarily attached to His essence than are other created things.* (To

> Mersenne, 27 May 1630, AT 1:152–53, CSMK3 p. 25 – my emphasis)

According to this passage, God has voluntary creative control over the essences of things, and those essences somehow determine the eternal truths. Thus since God has voluntary control over the essences of the objects He creates, He has voluntary control over the eternal truths.[3] From this consideration, Descartes is led to the creation doctrine in virtue of some of his other, sometimes idiosyncratic, theological views: in particular, he holds a version of the Divine Simplicity thesis according to which God's will and intellect are the same faculty. God's believing a proposition does not, like for us, involve a movement in God's will following a light in God's intellect. Rather, they, being the same divine faculty, move in unison. Thus God's will is not constrained by considerations of what is necessary, good, or true. As Frankfurt puts it, "there are no truths prior to God's creation of them, His creative will cannot be determined or even moved by any considerations of value or rationality whatever." (Frankfurt 1977, p. 41). This remarkable view is espoused and defended in the replies to the sixth objections to the *Meditations*, which I quote at length:

> As for the freedom of the will, the way in which it exists in God is quite different from the way in which it exists in us. It is self-contradictory to suppose that the will of God was not indifferent from eternity with respect to everything which has happened or will ever happen; for it is impossible to imagine that anything is thought of in the divine intellect as good or true, or worthy of belief or action or omission, prior to the decision of the divine will to make it so. I am not speaking here of temporal priority: I mean that there is not even any priority of order, or nature, or of 'rationally determined reason' as they call it, such that God's idea of the good impelled Him to choose one thing rather than another. For example, God did not will the creation of the world in time

[3] Frankfurt (1977, p. 39) notes that the eternal truths concern not just the essences of existing things, but also the essences of things which God has not made to exist. Frankfurt suggests that such objects only have *objective existence* in the sense of being objects of awareness, but need not be really existing.

> because He saw it would be better this way than if He had created it from eternity; nor did He will that the three angles of a triangle should be equal to two right angles because He recognized that it could not be otherwise, and so on. On the contrary, it is because He willed to create the world in time that it is better this way than if He had created it from eternity; and it is because He willed that the three right angles of a triangle should necessarily equal to two right angles that this is true and cannot be otherwise…(Sixth Replies, AT 7:431–432, CSM2 p. 291)

This makes it clear that Descartes takes God's act of creation to be totally unconstrained by considerations of truth or even necessary truth, and that these truths are themselves dependent on God's free will.[4] Similar claims concerning the dependence of necessary truths on God's will are made in the following passage, from a few years after Descartes wrote the replies to the Sixth objection, which is noteworthy for the use of the law of non-contradiction as an example:

> I turn to the difficulty of conceiving how God would have been acting freely and indifferently if He had made it false that the three angles of a triangle were equal to two right angles, or in general that contradictories could not be true together. It is easy to dispel this difficulty by considering that the power of God cannot have any limits, and that our mind is finite and so created as to be able to conceive as possible the things which God has wished to be in fact possible, but not be able to conceive as possible things which God could have made possible, but which He has nevertheless wished to make impossible. The first consideration shows us that God cannot have been determined to make it true that contradictories cannot be true together, and therefore that He could have done the opposite. The second consideration assures us that even if this be true, we should not try to

[4] What makes this view so remarkable is, perhaps, best brought out by keeping Leibniz in mind as a foil to Descartes, as the distinction between God's will and His intellect, and furthermore the fact that God is compelled to create the best possible world, is the key move in Leibniz's theodicy.

> comprehend it since our nature is incapable of doing so. (To Mesland, 2 May 1644, AT 4:118–119, CSMK3 p. 235)

The most important part for my purposes here is what is shown by the first consideration.[5] There is dispute as to how to read "therefore that He could have done the opposite," but on the simplest reading the opposite of making it true that contradictories cannot be true together is to make it that (some) contradictories *can* be true together.[6] From this reading of the passage, Frankfurt draws the following conclusion:

> God was free in creating the world to do anything, whether or not its description was logically coherent …Descartes evidently thinks that God could have omitted creating the essence 'circularity' entirely. In that case there would be *no* eternal truths about circles…Descartes also evidently thinks that God, while creating the essence 'circularity,' could have made it different from what we conceive it to be. In that case there would be eternal truths about circles, but they would differ from – and perhaps be the negations of – the propositions that are necessarily true of circularity as we now understand it. (Frankfurt 1977, p. 42–43).

So the interpretation offered by Frankfurt is as follows: God is the creator of essences, and essences determine the eternal truths. God could have failed to create some essences He did, hence making some of the eternal truths

[5] The second consideration is also important, as it raises the apparent possibility that what Descartes meant in these passages was *merely* that we are not in a position to judge the nature of God's power, and the freedom in which He created the world. This reading, which takes Descartes' point here to concern not what God's nature is like but rather what we have warrant to assert about it, would seem to make Descartes' claims here easier to come to grips with. I think, however, that while the 1644 Letter to Mesland supports this 'deflated' reading, the sixth replies, as well as a number of the other texts propounding CD, are more naturally interpreted as making claims about metaphysical reality itself, rather than about our ability to conceive it. For this reason, I'll proceed, as I have, reading the texts in this metaphysically loaded way.

[6] For comparison, see Ishiguro's (1986) for her alternative reading of 'doing the opposite' in light of her account of Descartes on negation.

false. Furthermore, God could have made some essences to be different than we understand them to be, even in ways which are logically incoherent.

How are we to understand the 'could' in the passages above? Clearly it can't pick out logical possibility, as we have understood Descartes to claim that God could bring about the logically impossible. Frankfurt's proposal, put briefly, is that this "could" expresses metaphysical possibility, whereas the necessity of the eternal truths is merely *epistemic*. Eternal truths only *appear* to be metaphysically necessary to us because God has created our minds in such a way that we can't conceive of their falsity. To return to the problem as stated in §1, Frankfurt reads (1) as a fact about our minds – i.e. eternal truths only appear necessary to us given the nature of our minds – and (2) as a fact about what it is possible for God to bring about (see (Frankfurt 1977, p. 45) for details).

As a final point, Frankfurt considers a proposal, for which he cites Gueroult (1953), that God has unlimited control only over essences which are not His own. So, Gueroult claims, there are some genuinely necessary truths: namely, truths about God's essence, such as that God is omnipotent and is not a deceiver (Gueroult lists more claims, but these are the ones which matter for my purposes). Frankfurt rejects this, pointing to some passages in which Descartes seems to claim that God's powers are *fully limitless*, or at least that this the position we must take, in light of our epistemic position:

> For my part, I know that my intellect is finite and God's power is infinite, and so I set no bounds to it; I consider only what I can conceive and what I cannot conceive, and I take great pains that my judgment should accord with my understanding. And so I boldly assert that God can do everything which I conceive to be possible, but I am not so bold as to deny that He can do whatever conflicts with my understanding – I merely say that it involves a contradiction. (To More, 5 Feb. 1649, AT 5:272, CSMK3 p. 363)

So Frankfurt solves the apparent contradiction between (1) and (2) by treating (1) as weaker than it first appears. Rather than using metaphysical necessity and so contradicting (2), (1) uses a kind of epistemic necessity, so there is only an apparent contradiction between them. There are a number of criticisms of universal possibilism, but the most substantial is that his weakening of (1) is a bad fit for Descartes' broader philosophical commitments. Curley's arguments to this effect are quite persuasive, and it is to them, and his positive proposal, that I now turn.

2.2. Curley: Limited Possibilism

Curley (1984) develops his account, *limited possibilism*, in response to a suggestion from Geach (1973), according to which while the eternal truths are necessarily true, they are not *necessarily* necessarily true. So Curley's view accepts (1), but replaces (2) with the weaker:

(2′) Eternal truths are possibly possibly false.

This dissolves the apparent contradiction in CD so long as Descartes is not committed to the distinctive principle of the modal logic S4, as mentioned above. So while Frankfurt seeks to weaken (1) to preserve coherence, Curley seeks instead to weaken (2). The key motivation for this move is that Frankfurt's weakening of (1) fails to do justice to the role played by eternal truths in Descartes' philosophy apart from concerns with CD. When discussing truths dependent on true and immutable natures in the *Meditations*, for instance, Descartes seems to commit himself to their necessity, and it seems that this necessity is significant for his broader project. Curley writes:

> Consider the ontological argument. As Descartes expounds this, it requires the assumption that I conceive of countless things which have true, immutable and eternal natures, even though they may never have existed or have been thought of (*Fifth Meditation*, AT 7:64, CSM2 p. 44). These eternal natures do not depend on my mind; my thought does not impose any necessity on things, rather the necessity of the things themselves determines me to think of them in the way that I do…Moreover, not only do we perceive that the truths of mathematics are necessary, sometimes, at least, we perceive clearly and distinctly that they are necessary. If they aren't in fact necessary, then it looks as though Descartes will have to give up his criterion of truth. Not everything we perceive clearly and distinctly is true. (Curley 1984, p. 572)

So Descartes is committed not just to the conceptual or epistemic necessity of eternal truths, but to their metaphysical necessity. Some of Descartes' discussion surrounding CD indicates not only that he thought that God could have willed the eternal truths false, but that He did, in fact, will them to be necessarily true. Consider the following passage, in which Descartes responds to an objection regarding the immutability of the natures in the *Meditations*:

> You say that you think it is 'very hard' to propose that there is anything immutable and eternal apart from God. You would be right to think this if I was talking about existing things, or if I was proposing something as immutable in the sense that its immutability was independent of God. But just as the poets suppose that the Fates were originally established by Jupiter, but that after they were established he bound himself to abide by them, so I do not think that the essences of things, and the mathematical truths which we can know concerning them, are independent of God. Nevertheless I do think that they are immutable and eternal, since the will and decree of God willed and decreed that they should be so. (Replies to the Fifth Objections, AT 7:380, CSM2, p. 261)

So there is substantial evidence that Descartes is committed to (1) with 'necessity' understood in a robustly metaphysical sense. Limited possibilism, by contrast, provides a way for a truth to be both freely created and necessary in this robust sense. On this line, while eternal truths are metaphysically necessarily, God was not necessitated to will them to be necessarily true, and it is this latter claim which expresses God's freedom in creation (Curley 1984, pp. 579–581). Limited possibilism dissolves the apparent contradiction by turning on this difference between God's willing a proposition to be necessary and His being necessitated to will it, which distinction Curley expresses in terms of *iterated modalities*:

> [The] suggestion is that we should understand Descartes's doctrine of the creation of the eternal truths as involving, not a denial that there are necessary truths, but a denial that those which are necessary are necessarily necessary…Descartes wants to allow that there are some propositions which are in fact impossible, but which might have been possible, and others that are in fact

> necessary, but might, nevertheless, not have been necessary. There is nothing epistemic about these ‘mights.’ We are not saying: ‘These things *seem* necessary, but, for all we know they might not *be* necessary.’ We are saying: ‘These things *are* necessary, but there is nothing necessary about *that*.’ (Curley 1984, p. 581–583).

Limited possibilism can be stated in very simple terms:

(LP) For every proposition ϕ, ϕ is possibly possible.

Curley presents a positive argument for this thesis, using a natural deduction system and some modal premises which Descartes plausibly accepts.

Limited possibilism has been criticized by a number of commentators. McFetridge (1990, pp. 179–180) and Van Cleve (1994) both use a variation on Curley’s formal argument to prove *universal* possibilism also follows from similar premises. Van Cleve takes this to prove that Descartes’ metaphysics is inconsistent, whereas McFetridge just notes that it puts pressure on Curley’s view. Beside noting this logical property of Curley’s proposal, the general tenor of the criticisms of limited possibilism is that (2′) does not adequately capture Descartes’ commitments concerning CD, or that it somehow renders God’s powers *too weak*. A general version of this criticism against limited possibilism had been put forward by Plantinga (1980) before the publication of Curley’s paper, and versions have since been presented by Alanen (1985), Bennett (1994), Kaufman (2002), and recently by Saint-Germier (2018). Curley admits (Curley 1984, p. 590) that limited possibilism, as a modal theory, is not substantially more plausible than universal possibilism, but he seems to hold that something like limited possibilism is needed in order to render CD coherent, noting that CD “faces severe difficulties, even on the most charitable of interpretations.” (Curley 1984, p. 597).

2.3. McFetridge and Saint-Germier: Bimodalism

The general trouble with CD is that (1) and (2) cannot be coherently captured with one pair of modal operators which are connected by modal duality – i.e. that ‘ϕ is necessary’ is equivalent to ‘it’s not possible that not ϕ’ and ‘ϕ is possible’ to ‘it’s not necessary that not ϕ.’ So if you understand

Descartes as countenancing only one kind of metaphysical modality, then (1) or (2) has to give. Frankfurt renders (1) in terms of epistemic necessity in order to make room for the possible falsity of the eternal truths, whereas Curley replaces (2) with (2′) in order to make room for their metaphysical necessity. The most natural solution calls, rather than rejecting one of these, for a disambiguation of the terms occurring therein.

McFetridge, in the posthumously published (McFetridge 1990), defends a bimodal view of Descartes' modal metaphysics. McFetridge notes a deep ambivalence in Descartes' writing between whether eternal truths depend on our minds or not, which point has recently been reinforced by De Rosa (2011). Furthermore, he takes this ambivalence to be reflected in the two proposed solutions discussed previously. In short, universal possibilism seems plausible in light of the tendency in some of Descartes' writings to treat the necessity of the eternal truth as somehow epistemic, or determined by human minds and thinking, while limited possibilism seems plausible in light of the tendency in other of Descartes' writings to treat that necessity as mind independent, and hence more like alethic necessity. McFetridge invokes a bimodal solution:

> In one sense, 'necessary$_1$,' certain propositions are necessarily$_1$ true (though not necessarily necessary$_1$). In another sense, 'necessary$_2$,' no truths are necessarily$_2$ true....Necessity in one sense would be, in another sense would not be, 'a function of the structure of our mind.' (McFetridge, 1990, pp. 180–181).

This account provides the grist for a simple solution to the apparent contradiction between (1) and (2), as sketched in the introduction. The trick is just to allow that while some propositions are necessary$_1$, this does not entail that they are necessary$_2$. If we allow that the eternal truths are necessarily1 true and possibly$_2$ false, one can provide for a consistent disambiguation of the central passages concerning CD.

In fact, McFetridge builds in more substantial assumptions about the relative behaviour of his modalities. In particular, he assumes that no propositions are necessary$_2$, relying on passages, and some of the reasons, Frankfurt appealed to. He also singles out an instance of this, namely that no proposition is necessarily$_2$ necessary$_1$, as part of his explanation of the passages relied on by Curley, according to which while God willed the eternal truths to be necessary (necessary$_1$), God was not necessitated (necessary$_2$) to will this. Finally and I'll come back to this point in §5, he argues that the set of possible1 propositions must be identical to the set of conceivable propositions – i.e. these must be coextensive (McFetridge 1990, pp. 191–194). The reasoning in this passage, while I'll come to, seems to provide a fatal objection to the kind of epistemic/alethic bimodalism proposed by Frankfurt.

Saint-Germier (2018), working in McFetridge's bimodal framework, delves further into issues concerning the relationship between conceivability and possibility, some of which we'll come back to in §5. In addition, he (Saint-Germier 2018, pp. 4807–4811) argues, against Alanen (1985) and Kaufman (2002), that the modal-ity which expresses facts about what God could have done (McFetridge's "necessary$_1$") is a genuine modal notion, and so I refer the interested reader to that discussion in Saint-Germier's work.

The bimodal approach seems to be the best way to dissolve the apparent contradiction in CD, and the disambiguation strategy allows one to retain the important insights of universal and limited possibilism without falling into (the most obvious) pitfalls of either theory. Furthermore, McFetridge and Saint-Germier fill in the structure of such an account in some ways that allow us to develop important insights into the interplay of various modal notions at work in Descartes' philosophical work. There is, however, an avenue of improvement on this kind of bimodal theory, which I'll seek to follow, and a point of disagreement with McFetridge and Saint-Germier which falls out of the approach I follow.

The avenue of improvement concerns a point which McFetridge raises be-fore going into the structure of his theory (McFetridge 1990, p. 175–176). He notes that CD commits Descartes to the view that the ground of the necessity$_1$ of necessary$_1$ truths seems to be contingent$_2$. That is, the ground, or explanation, for why an eternal truth is necessary$_1$ is that God willed it to be – but since God's will is free, God could have done otherwise. McFetridge notes this interesting aspect of Descartes' metaphysical picture, and draws some consequences, but does not say much about how necessary truths are grounded besides making an oblique reference to God's will. Saint-Germier also does not aim to provide a "full conceptual analysis of modality" (Saint-Germier 2018, p. 4809) for Descartes. That is, neither author seeks to provide an account explaining *why eternal truths are necessary* (in whatever sense they are necessary). It would be an improvement on existing work in the bimodal interpretation of CD to provide an account which grounds/explains the necessity of the two kinds of necessary truths in appropriately Cartesian terms. One may respond that the weaker form of necessity (McFetridge's necessary$_1$) can be grounded in God's act of creation. This is surely correct, but it seems to me that more can be said about how this grounding works, and that saying more would lead further plausibility to the account. In the rest of the paper, I'll argue that the best story appeals to Descartes' *essentialism.*

Before moving on to this, I'll preface the point of disagreement. Both McFetridge and Saint-Germier follow the universal possibilist line in that they hold the the stronger notion, necessity$_2$, applies to no propositions. While I can see the reasons for adopting this, a bimodal account need not do so, and the biessentialist line I propose gives some reasons to think that there are some propositions which Descartes should take to be necessary even in the stronger sense. Namely, following Gueroult, propositions about God's nature are necessary in both senses. I'll discuss this point further in the next, and the last, section of the paper.

3. Essences and Two Modalities

I seek to argue that the bimodal disambiguation fits well with Descartes' broader metaphysical commitments, and I seek to do so in terms of attributing to him a species of essentialism. According to essentialism, as I use the term, modal truths are to be understood as reducible to essences, or collections thereof, according to (something like) the following schema:

- ϕ is necessary iff ϕ is made true by the salient essence(s).
- ϕ is possible iff ϕ is not made false by the salient essence(s).

For simplicity, I'll occasionally express the definiens of 'ϕ is possible' by saying that (the truth of) ϕ is *compatible* with the salient essence(s). Furthermore, I'll assume that any proposition is either made true by some essence or made false by some essence, so that ϕ is possible iff it is not necessarily false, and similarly that ϕ is necessary iff it is not possibly false.

3.1. What Are Cartesian Essences?

The kind of essentialism I attribute to Descartes here bears a notable resemblance to the essentialism which Embry attributes to Suárez. (Embry 2017). Embry argues that, in disputation 31 of *Disputationes Metaphysicae* (see (Suárez 1983, pp. 178–211), Suárez rejects a brand of essentialism developed by Henry of Ghent, according to which the essences which ground eternal truths in essences which "necessarily have essential being from eternity" (Embry 2017, p. 559), and replaces it with a view according to which essences, while grounding eternal truths, may be non-existent, enjoying only a kind of potential being (Embry 2017, pp. 558–561). Embry's account relies on sutble features of Suárez's metaphysics which need not concern us here, but the claim that non-existent (or merely potentially existent) essences ground necessary truths seems appropriately attributable to Descartes as well.[7]

The core of the account of Cartesian essences involves a few claims, for which well-known texts provide evidence:

(a) The properties of essences are not determined by what we are able to conceive (see *Fifth Meditation, AT 7:64, CSM2 p. 44–45*, *Conversation with Burman* AT 5:160, CSMK3 p. 343).

(b) Essences of created things are freely created by God (Letter to Mersenne, 27 May 1630, quoted above).

(c) Essences ground (in some way) eternal truths (same as for (b))

Point (c) is needed in order to make grounding modal properties of propositions in properties of essences associated to those propositions (as I do here) plausible. There are a number of potential ways to think about the relationship between essences and eternal truths that's consistent with Descartes' commitments, the strongest of which holds essences and eternal

[7] Alanen, Cronin, and Pessin (Alanen 1991, Cronin 1960, Pessin 2010) situate Descartes' views surrounding CD in relation to his contemporaries and predecessors, and provide ample evidence of Suárez's influence on Descartes. This influence is mostly negative, for instance in that some of Descartes' examples in CD are clearly stated as the negations of some of Suárez's claims (see Cronin), but this, by itself, doesn't provide reason to think that Descartes was not influenced by Suarez's essentialism, in broad strokes. This is the position I'll defend.

truths to be *literally identical*, which get to if we take the 27 May 1630 letter to Mersenne at face value. For my purposes, all that's needed is that Descartes takes eternal truths to be true *in virtue of* essences, and this certainly is supported by that letter to Mersenne even if we don't attribute to Descartes this strong identity thesis.

Given that, point (b) indicates why God has free control over eternal truths – at least those which concern created things. For since God has free control over essences, and essences ground those eternal truths, then God has free control over the eternal truths. Finally, point (a) provides the grist for rejecting epistemic/alethic theories, that it is incorrect to think of the necessity of the eternal truths in purely epistemic, or conceptual, terms. This way, the essences are able to ground modalities which are genuinely metaphysical, insofar as they reflect an aspect of reality, rather than being reliant on what are able to conceive.

The issue of providing a clear and coherent account of Cartesian essences, and their ontological status, is itself a thorny interpretive problem. As De Rosa (2011) argues, there is a deep tension in Descartes' writing pulling between *Platonist* accounts of essences, which situate them in God's mind or decrees, and *conceptualist* accounts, which situate them in the human mind. She claims that this tension requires a deep rethink of the problem space, and the case for this is compelling.[8]

I am not able to resolve the difficulties raised by De Rosa here. I take it as plausible that Descartes is an essentialist in the sense I propose above, while admitting that there are difficulties with giving a compelling story

[8] A broadly Platonist line is a better fit for my proposal, though I would, following Kenny (1968, pp. 155–156), emphasize the *Meinongian* (or, perhaps better, *Noneist*, in the sense of (Routley 2018, Priest 2016) aspect of Descartes' thought, according to which something may be non-existent while still, it would seem, potentially being the subject of our ideas and the grounds for the truth of propositions. It seems likely to me that a kind of Platonism which appealed to this Meinongian aspect may not contradict the apparent conceptualism of the *Principles*, in that it might allow for essences to be non-existent objects, except for the extent to which they can come to exist in our minds. This way, the story could go, essences, if they ever exist, exist in some mind, but that they needn't have this existence in order to have their essential properties, and so their having those properties is not determined by whether they do, in fact, exist in some mind.

about what exactly Cartesian essences are. For my purposes, the reduction of modal facts to essences is of the most importance, and, following Frankfurt, I take it that the cited texts above provide enough evidence to render this claim plausible.[9] Having said this, in order to argue that the essentialist reading does provide for grounds for alethic modalities, rather than epistemic modalities, the account I need is broadly Platonist in character. Perhaps the most plausibly such account is Rozemond's *moderate platonism* (Rozemond 2008), according to which essences are dependent on God (and hence God's mind), but are independent of our minds. I won't go into more detail defending the use of this account, but will rather take it for granted.

3.2. Distinguishing Inner and Outer Modalities

There are two salient (collections of) essences which, substituted into the above schema, provide definitions of modalities which we should understand Descartes as employing. These are, first, God's essence and, second, the essences of created things. The latter class is invoked explicitly in explaining and justifying (2) in the letter to Mersenne (27 May 1630) cited above. The schema with the essences of created things defines what I'll call the *inner* or *i*-modalities (these are analogous to McFetridge's 1-modalities):

- ϕ is i-necessary iff ϕ is made true by the essences of created things.
- ϕ is i-possible iff ϕ is not made false by the essences of created things.

As an example, it is i-necessary that triangles have three angles summing to two right angles because this proposition is made true by the essence 'triangularity'; nothing can have that essence without having three angles with this property. Furthermore, the proposition 'some triangle has a right angle' is i-possible in virtue of the fact that this property is compatible with an object being triangular (since there are right triangles), but 'some

[9] Secada (2000) discusses a different sense in which Descartes is an essentialist, namely in that Descartes holds that knowledge of the essence of something precedes knowledge of its existence. While not of direct import to my work here, Secada's book provides interesting discussion of related topics, including the late Scholastic context of Descartes' metaphysics.

triangle has two right angles' is not, because the property 'has two right angles' is ruled out by being triangular.[10]

The other salient essence here is God's. Descartes does appeal to facts about God's nature in his philosophical work. For instance, Descartes claims that God's essence includes existence in §14 of Part 1 of the *Principles*. The same point is, more famously, made in the Fifth Meditation (AT 7:65, CSM2 p. 46), where it is claimed "existence can no more be separated from the essences of God than the fact that its three angles equal two right angles can be separated from the essence of a triangle." Here the point seems to be that we recognise existence as belonging to the essence of God in precisely the same way as we recognise properties belonging to the essences of mathematical objects.

God's essence provides the grist of what I'll call the *outer* or *o*-modalities (analogous to McFetridge's 2-modalities):

- ϕ is o-necessary iff ϕ is made true by God's essence.
- ϕ is o-possible iff ϕ is not made false by God's essence.[11]

The solution to the problem with CD then relies on the claim that, for Descartes, there are many truths which, though they are guaranteed to hold in virtue of the essences of created things, both they, and their contradictories, are compatible with God's essence. While such truths are i-necessary, they are o-contingent, or *merely* o-possible. The import of CD is that every eternal truth not concerning God is like this. Expanding the example from before, while 'triangles have three angles summing to two right angles' is made true by the essence *triangularity*, it is not made true by God's essence. God's essence is compatible with it, so it's o-possible, but God's essence is also compatible with the opposite, 'triangles have three angles summing to something other than two right angles,' so this is also o-possible. While God created the essence *triangularity*, His essence does not ground its essential

[10] With Descartes, I consider only *Euclidean* geometry.

[11] It is highly plausible that 'ϕ is o-necessary' implies 'ϕ is i-necessary' and that 'ϕ is i-possible' implies 'ϕ is o-possible,' so the truth conditions for the i-modalities should, strictly speaking, be stated in terms of 'the essences of created things or God's essence' rather than merely 'the essences of created things.' The truth conditions are stated in the slightly less correct manner for simplicity.

properties, and hence is compatible with it having other properties, or perhaps with there being no such essence.

With this picture, the apparent contradiction between (1) and (2) is resolved by reading them as follows:

(1) Eternal truths are i-necessarily true.
(2) Eternal truths are o-possibly false.

The eternal truths are necessary, but only in the sense that they are made true by the essences created by God, not in the sense that they are made true by God's essence itself. Furthermore, and this comes out as the radical part of Descartes' doctrine, God's essence is compatible with the falsity of the eternal truths, and even with the truth of propositions contradicting the eternal truths. This is the sense in which God was free to make it not true that all the radii of the circle are equal (To Mersenne, 27 May 1630) and that in which He "cannot have been determined to make it true that contradictories cannot be true together, and therefore that He could have done the opposite" (To Mesland, 2 May 1644). God's essence neither makes it true that all the radii of the circle are equal nor does it make it true that contradictories cannot be true together. The eternal truths associated with these things are all i-necessary, but o-contingent, and furthermore this is explained by appeal to the explanation of the modal properties of these propositions.

Bimodalism also explains the distinction, relied on by Curley, between God's making ϕ necessarily true and God's being necessitated to make ϕ true. In my framework, "God makes ϕ necessary" expresses the fact that ϕ is i-necessary, i.e. that God makes ϕ true by making the essences of creating things in the appropriate way. On the other hand, "God is necessitated to make ϕ true" expresses the claim that ϕ is o-necessary, so that it's truth is grounded in God's own nature. On this way, the difference between the iterated modality reading and the bimodal reading can be cashed out in terms of the grounds of the necessities in question, rather than just in terms of the logical form of the central claims.

One more important point to note here is that on this reading not *every* eternal truth should be understood as *merely* o-possible, as God's essence does make some claims true. Examples of such claims which play an central role in Descartes' broader philosophical project are 'God exists' and 'God is not a deceiver,' but there are many such claims which are plausible candidates: 'There is one unique God'; 'God has every perfection'; 'God is immutable.' Frankfurt, as well as Saint-Germier and McFetridge, hold that God's limitless power should be expressed in universal possibilist terms, according to which no proposition is necessary ('o-necessary' or 'necessary$_1$' for McFetridge and Saint-Germier). I differ in taking it that, since God's nature does ground some propositions, those propositions are o-necessary. One might worry that this winds up putting inappropriate *limits* on God's power. In response to this, I have a couple points.

The first is that, following Kaufman (2002, pp. 38–39), it is plausible to understand Descartes as committed to the view that something only counts as a *limit on God* if it comes from outside God. For instance, in the 27 May 1630 letter to Mersenne, Descartes is concerned to argue that eternal truths are under God's control *because* He is their author, and furthermore that these are not necessarily attached to his essence. (AT 1:152–153, CSMK3 p. 25) Furthermore in the Replies to the Sixth Objections, Descartes is concerned to indicate that the eternal truths are necessary only *because* God willed it, and that "nothing impelled Him to choose one thing rather than another." (AT 7:431–432, CSM2 p. 291) Finally, Descartes, in the Replies to the Fifth Objections, seems to indicate that God's will could be bound by previous decrees (in the analogy with Jupiter, the Styx, and the Fates – see Kaufman for discussion of this passage). This suggests that Descartes' main concern in not putting limits to God's power is to avoid positing something outside God which limits Him. This leaves room for the claim that God's indifference and freedom is compatible with constraints which come from God Himself. Since God's nature is not external to Him, constraints imposed by God having the nature He does are compatible with God's freedom, as Descartes is most concerned to defend it.[12] The creation can still be understood as free even in spite of the fact that God's own nature is not within His control in the same way that the natures of mathematical objects are.

[12] The letter to More 5 Feb. 1649 does see Descartes refusing to set bounds to God's power, but in light of this way of understanding the kind of freedom in question, it's not obvious that we have to read this passage in such a way as to motivate universal o-possibilism.

The second point is just to note that Descartes holds that, in contradistinction to the essences of created things, God is not His own efficient cause. In the replies to the fourth objections, Descartes writes “the phrase ‘his own cause’ [applied to God] cannot possibly be taken to mean an efficient cause; it simply means that the inexhaustible power of God is the cause or reason for his not needing a cause” (AT 7:236, CSM2 p. 165). The main reason, as I understand it, for taking (most) eternal truths to be o-possibly false is that God, in freely creating these things (as “their efficient and total cause,” Letter to Mersenne 27 May 1630), could have made them some other way. Since Descartes denies that God stands in this causal relationship to Himself (and presumably His essence), the main positive reason for taking these propositions to be o-possibly false is undercut. This provides a textual reason to deny that Descartes holds that God’s nature is under His control in the same way that the natures of mathematical objects are.

As a final point against universal possibilism, note that none of the passages cited so far (and these comprise most of Descartes’ writing on CD) explicitly concern propositions concerning God Himself. Descartes could have, following the 1644 letter to Mesland, claimed somewhere that “God cannot have been determined to make it true that He exists, and therefore He could have done the opposite.” The example Descartes chooses in this passage, the law of non-contradiction, is striking, but it’s not as striking as this. The other examples concern mathematical entities, and even whether there is a world at all (27 May 1630 letter to Mersenne), but none concern God Himself, or His nature. Given what Descartes does include in the scope of CD (almost every proposition), the fact that he didn’t take the one further step to consider examples about God is some evidence that the view was not intended to stretch *that* far.

I would like to caution against being misled by my, and Descartes,’ occasional use of ‘God could’ locutions in order to express o-possibilities. For instance, when discussing CD, Descartes considers (in the 2 May 1644 letter to Mesland) those ‘things which God could have made possible, but which He has nevertheless wished to make impossible.’ On my reading, ‘God could have made ϕ true’ is just a way of expressing ‘ϕ is o-possible.’ In particular, it should be noted that ‘God could have made ϕ true’ does not entail the i-possibility or o-possibility of a change in God’s will. Indeed, since God is immutable, it would seem to be o-impossible for God's will to change. God's freedom to create the world other than it was created (or to not create the world at all, as per the letter to Mersenne, 27 May 1630) is not a matter of the possibility (of either kind) of God's will changing, but rather is just a matter of the falsity of propositions expressing those circumstances being compatible with God's essence.

This suffices to express the core part of my proposal, according to which Descartes' essentialism about modal facts, and consideration of two very different kinds of essence, leads him to adopt a bimodal metaphysics in virtue of which the apparent inconsistency in CD is rendered as *merely* apparent. This presentation of the biessentialist interpretation leaves many questions unanswered, but I take it that the essentialism is, in broad strokes, plausible as a reading of Descartes' modal metaphysics, and that in virtue of this the bimodalism is rendered more satisfying as a solvent for the apparent inconsistency in CD. While I can't hope to address every question concerning biessentialism and how it interacts with Descartes' myriad commitments, in the rest of the paper I will address two natural questions. The first concerns how we can understand logical truths to be grounded in essences, for Descartes, and the second concerns an upshot for biessentialism coming from the Fourth Meditation discussion of the truth rule.

4. Universal and Logical Truths

The examples of necessary truth I've discussed the most thus far are those concerning particular essences such as *tringularity*. It is easy to see how one might hold that necessary truths concerning triangles can be explained by appeal to the feature of this essence. Descartes, however, also considers the law of non-contradiction (LNC) as in the scope of CD. LNC is one of those propositions which, while necessary, is nonetheless such that God could have made it false. This example would seem to have broader import, insofar as its inclusion suggests that Descartes would consider all logical truths as in the scope of CD. In addition to logical truths, there are a number of plausible candidates for CD which concern properties which apply universally, such as "things that are the same as a third thing are the same as each other" (*Regulae* Rule 12) and "what is done cannot be undone"

(*Principles* Part 1, §49). How should we understand the necessity of these universal truths as grounded in the essences of created things?

It is reasonably plausible to hold, following Fine (1994) and Hale (2018), that eternal truths which hold regardless of the object in question (but which still concern properties of *objects*, and do not concern the truth of propositions, or other more clearly logical properties of propositions) are grounded in *collections* of essences of created things. The ground of the necessity of "the whole is greater than the part" can be explained in terms of the essences of any object which can have, or does have, parts. Then the necessity of this eternal truth is universal in being grounded not just in some essence or other, but in any essence of the appropriate kind of object. For those eternal truths which apply to every object whatever, the natural grounds of i-necessity would be the collection of all essences. For example, the necessity of the claim "everything is self-identical" is plausibly grounded in the necessity of the instances, each of which is grounded in the essence of the object in question. This kind of approach to grounding logical necessity has recently been suggested by Shalkowski (2004, p. 79) who suggests that "logical necessities might be explained as those propositions true in virtue of the natures of every situation or every object and property, thus preserving the idea that logic is the most general science." It is, in any case, a plausibly essentialist explanation of the grounding of universal i-necessary truths in the essences of particular created things, and I don't see any reason to think it's incompatible with Descartes' commitments. Having said that, it's not obviously the right approach to take for logically necessary truths which seem to concern not objects but propositions, and their truth and falsity. Since LNC falls under this description, it seems that a different story should be given for such i-necessities.

I've mentioned Shalkowski, but in addition to him a number of contemporary essentialists have sought to provide accounts of how logical truths, and their necessity, can be explained by appeal to essential properties. Fine (1994) claimed that logical necessities are true/necessary in virtue of the natures of *logical concepts.*[13] This kind of account leaves some flexibility for

[13] This fits into Fine's broader essentialism, according to which all necessities belonging to a particular domain or discipline (e.g. conceptual, logical, physical ...necessities) "are true in virtue of the characteristic concepts and objects of the

what the salient logical concepts are. According to one proposal, due to Correia (2012), these are *inference rules* in a natural deduction system; Hale proposes that certain *logical functions*, such as, perhaps, truth functions, provide the grounding; and Shalkowski (2004), in passing, suggests *truth bearers* or *propositions* themselves.

These are interesting accounts, but the difficulty with trying to read Descartes as potentially accepting any of them would involve restating them without appeal to the machinery of contemporary logic. This would seem to rule out an approach like Hale's, which builds in the apparatus of contemporary model theory. Having said that, both Correia's and Shalkowski's suggestions seem able to stand even if the contemporary machinery is removed. Furthermore, it seems that one could find room for something like either in Descartes' broader philosophical work. So I'll consider these two kinds of account in turn, and suggest how one could make Cartesian sense of them. I won't come down on one or the other as the best view, because I don't see overwhelming textual reasons in favour of either, and furthermore suspect that such reasons will not be forthcoming. Descartes' explicit discussions of logic, as studied by his predecessors and contemporaries, are few and mostly concern his unhappiness with the subject – see Gaukroger (1989), especially Chapters 1 and 2.[14] This suggests that we'd be hard pressed to give a robust defense of any positive account of the grounding of logical necessities. So I'll aim only to sketch two accounts (which are not, as far as I can see, mutually incompatible), and claim that these indicate that an appropriately Cartesian essentialist reduction of logical necessities is possible.

4.1. Grounding Logical Necessities in Inference

Correia's account provides an interesting inroad in that it can be understood in terms of *inference* (Correia 2012, pp. 646–650). For Correia, logical

discipline," while metaphysically necessary truths as those "which are true in virtue of the nature of all objects whatever." (Fine 1004, p. 9–10)

[14] It is worth noting, briefly, that while Descartes has little to say of logic that is kind, he does, in the Conversation with Burman, note that at least one of his complaints "applies not so much to logic, which provides demonstrative proofs on all subjects, but to dialectic, which teaches us how to hold forth on all subjects." (Conversation with Burman, AT 5:175, CSMK3 p. 350)

necessities are true in virtue of the correctness of certain modes of inference; for example, an instance of the law of excluded middle "ϕ or not ϕ," is logically necessary in virtue of the existence of a natural deduction proof of this proposition, from no premises, using the inference rules characterising disjunction and negation. The particulars of a natural deduction system are part of Correia's account, but it seems that we can get a reasonable account even if we remove this part. One could simply say that it is logically necessary that ψ sentence follows from ϕ in virtue of the correctness of inferring ψ from ϕ, without any claim that such an inference must follow the patterns of any proof system. Similarly, one could say that ϕ is logically necessary just in case it is always correct to infer ϕ. Since it is always correct to infer an instance of LNC, this is logically necessary. What's compelling about this kind of account is that Descartes, even while not generally concerned with the syllogistic, does concern himself with inference.

Gaukroger (Gaukroger 1989, Ch. 2) situates Descartes' conception of inference within his broader views concerning how we obtain scientific knowledge, in particular against the background of the *Regulae* account of such knowledge as obtained through two intellectual processes: *intuition* and *deduction* (see especially Rule 3, AT 10:366–370, CSM1 pp. 13–15). The former "consists in grasping one proposition or in grasping a necessary connection between two propositions, and it is equated with clear and distinct perception"(Gaukroger 1989, p. 50). In contrast, Descartes defines deduction as "the inference of something as following necessarily from some other propositions which are known with certainty," and he goes on to claim that "very many facts which are not self-evident are known with certainty, provided they are inferred from true and known principles through a continuous and uninterrupted movement of thought in which each individual proposition is clearly intuited." (*Regulae* Rule 3) So a deduction consists in a chain of propositions, some early ones of which are intuited, such that we intuit the necessary connection between those next to each other in the chain, so that we go on to intuit the latter propositions, including whatever the deduction is seeking to prove. At first, we go through a deduction with the use of memory, but the goal is to eventually be able to think through this chain "so quickly that we no longer have to rely on memory, with the result that we 'have the whole in intuition' before us at a single time" (Gaukroger 1989, p. 50).

In fact, the most concrete example Descartes gives in the *Regulae* concerns chains of propositions, each of which, being categorical, involves the comparison of *objects*. The example is "all A is B, all B is C, therefore all A is C." (*Regulae* Rule 14) He claims that the best way to think of this inference, and "all knowledge whatsoever" as involving comparisons between objects. He goes on to claim that such comparisons really do make up the bulk of our inferential activity, and discusses how we proceed from easy comparisons (where the objects share a nature) to more complex comparisons:

> the business of human reason consists almost entirely of preparing for this operation [that of comparing two or more things]. For when the operation is straightforward and simple, we have no need of a technique to help us intuit the truth which the comparison yields; all we need is the light of nature. We should note that comparisons are said to be simple and straightforward only when the thing sought and the initial data participate equally in a certain nature. The reason why preparation is required for other sorts of comparison is simply that the common nature in question is not present equally in both, but only by way of other relations or proportions which imply it. The chief part of human endeavour is simply to reduce these proportions to the point where an equality between what we are seeking and what we already know is clearly visible. (*Regulae* Rule 14, AT 10:440, CSM1 pp. 57–58)

The claim above concerning a certain nature, or a common nature between some things, is suggestive of one approach to grounding necessary connections between propositions, and hence grounding deduction and logical necessity. A necessary connection between propositions, of the sort to support deduction of one from the other, may be supported by there being things described in the two propositions which share an essence, and hence some essential properties. For instance, both triangles and circles share the essence *geometrical figure*, and so some inferences involving propositions discussing these two figures may be grounded in these shared properties (such as their having sides, angles, an area, etc…). Then the necessary connection is grounded in the shared essential properties of two objects, and this grounds the correctness of some associated inference.

It should be noted that this reading of the passage above involves an analogy, because "nature" as it occurs in the *Regulae* does not seem to mean the same thing as "essence," and its cognates, do in the rest of Descartes' writings, given that the former are primarily mental entities, and furthermore can be present to a greater or lesser extent in an object. However, the analogy here seems plausible to the following extent: when making a comparison between two objects, we can recognise some shared essential properties, and furthermore these properties can ground a necessary connection between the objects, and some propositions concerning them.

4.2. Grounding Logical Necessities in Truth Bearers

The account sketched above explains the fact that the conclusions of an inference follow necessarily from the premises, but it is less obviously able to provide a compelling explanation of the fact that certain claims, not directly about inference, seem to be logically necessary. Furthermore, the logical claim which Descartes explicitly considers is one of these.

On the face of it, LNC seems to be a claim about *truth* rather than about individual objects. This presents the difficulty in finding a plausible essentialist basis, but also suggests that we look to what grounds truth for Descartes, in order to discover what may ground the necessity of LNC. Unfortunately, the nature of truth (of propositions, not of objects) is another of topic which gets little discussion in Descartes' work. He seems to hold a version of the correspondence theory, as he claims "it is possible to explain the meaning of the word to someone who does not know the language, and tell him that the word 'truth,' in the strict sense, denotes the conformity of thought with its object." (Letter to Mersenne, 16 October 1639, AT 2:596–597, CSMK3 p. 139)[15] So, if we were to follow Shalkowski's suggestion and seek to ground logical necessity in truth bearers, then we should identify the latter as *thoughts*.

In particular, the kinds of thoughts which matter here "are as it were images of things" (*Meditations* 7:36, CSM2 p. 25), which he calls *ideas.* Given that thought is the principal attribute of mental substance (*Principles*

[15] Curley (1984, p. 572) notes the same passage as evidence that Descartes held a correspondence theory.

Part 1, §53), it's highly plausible that, for Descartes, individual ideas are either attributes or modes of that substance. Furthermore, the properties of individual ideas will be grounded in mental substance itself, and their necessary properties grounded in the essence, or attribute, of that substance. From this, the necessary properties of *truth*, since truth is a property of ideas, will be grounded in mental substance.

What LNC seems to claim is that it is not possible for a pair of contradictory ideas to conform to their object – they can't both be true. LNC, in stating a necessary property of ideas, is plausibly grounded in the ground of ideas. That is, it is a property of the nature of thought, the attribute of mental substance, that contradictory ideas cannot both be true. On this sketch, then, the necessity of logically necessary truths is grounded in the nature of the mind, rather than in the nature of any particular objects.

A natural worry is that such an account would wind up making logical necessities merely epistemic, rather than alethic, which is the kind of necessity I've claimed is enjoyed by eternal truths. I will just here note the difference between "ϕ is necessary in virtue of a thinking thing's being incapable of conceiving its falsity" and "ϕ is necessary in virtue of the nature of thinking beings." The former grounds the necessity of ϕ in *the fact that* we can't conceive its falsity, while the latter grounds this necessity in the nature of our mind (which itself determines what we are able to conceive). This difference is enough to indicate that the sketch above does render logically necessary truths as alethically necessary.

4.3. Logically Necessary Truths

The above sketches are merely sketches, and I doubt that more compelling stories are forthcoming from the text. I think, however, that these make it at least plausible that we can read Descartes as having an essentialist ground of logically necessary truths, however he understands these. Unfortunately, while he used LNC as an example in his discussion of CD, the rest of his work provides only hints as to how we should understand this – hence my goal has been to follow some hints, and to try to construct plausible stories from them.

5. God is O-Necessarily Not a Deceiver

An important question raised in §2.1 is that of delineating the scope of CD. Over just which necessary truths is God supposed to have free control? Put in terms of the vocabulary used here: which propositions are o-necessary? The point of contention here is whether God has free control even over eternal truths about Himself. On Frankfurt's line, God could have made it that He doesn't exist, that He is imperfect, etc…, since God could have made any truth false. Against Frankfurt, and in keeping Gueroult, my account has it that Descartes is committed to the claim that certain eternal truths about God are o-necessary. The reason for this is that some truths are made true by God's own essence, and hence are necessary in the strongest sense available. Examples of such truths are those ascribing essential properties to God such as existence, omniscience, eternality, and immutability. Descartes ascribes some of these properties to God, giving a list in the Third Meditation (AT 7:40, CSM2 p. 28), and furthermore more in the Fourth Meditation (AT 7:57, CSM2 p. 40) he ascribes certain perfections not just to God, but to God's essence: the meditator, after considering their own faculty of understanding, says "I at once form an idea of an understanding which is much greater – indeed supremely great and infinite; and from the very fact that I can form an idea of it, I perceive that it belongs to the nature of God." This suggests that omniscience, along with the other superlative properties Descartes ascribes to God in the previous meditation, should be understood as being ascribed to God's essence. This is in addition to Descartes' claims, in giving his ontological argument, that God's essence includes existence.

So on the biessentialist picture, there are a number of o-necessary propositions. I've already argued that this fact should not count against the picture, because we should not think of these propositions as *limiting* God, and in this section I'll provide an argument (apparently presaged by Spinoza) that "God is not a deceiver" is one of these. That this claim is o-necessary can already be seen by Descartes' claim that a will to deceive indicates an imperfection (AT 7:53, CSM2 p.37), and so is incompatible with the nature of a being, like God, which includes every perfection. There is, however, a further argument which is interesting in tying back to the Fourth Meditation more directly.

This point involves consideration of some finer points in Descartes' discussion of conceivability and possibility. McFetridge (1990) argues that Descartes is committed to the claim that ϕ is conceivable iff ϕ is possible (i-possible in the framework developed here). His defense of this interpretive claim relies on a passage in the replies to the second set of objections to the Meditations.[16]

> *If by 'possible' you mean what everyone commonly means, namely, 'whatever does not conflict with our human concepts,'* then it is manifest that the nature of God, as I have described it, is possible in this sense, since I supposed it to contain only what, according to our clear and distinct perceptions, must belong to it; and hence it cannot conflict with our concepts. Alternately, *you may well be imagining some other kind of possibility which relates to the object itself; but unless this matches the first sort of possibility it can never be known by the human intellect*, and so it does not so much support a denial of God's nature and existence as serve to undermine every other item of human knowledge. (Second Replies AT 7:150-151, CSM2 p. 107 – my emphasis)

The upshot of McFetridge's reading of this passage is that Descartes is committed to the coextensivity of conceivability and i-possibility. In addition, there is good reason to suppose that Descartes is committed to the fact that this equivalence is not only true, but o-necessary. The argument proceeds by considering an i-possible world, and a brief remark is in order

[16] As some background, Descartes is responding to Mersenne's criticism of Descartes' argument in the fifth meditation that God exists. Mersenne holds that what Descartes shows in that meditation is only that existence belongs to the nature of God, but not that God exists because Descartes has not successfully argued that God's nature is possible. (CSM2 p. 91) Descartes takes Mersenne to open the argument with the major premise "That which we clearly understand to belong to the nature of something can be truly asserted to belong to its nature." (CSM2 p. 106) In the following passage, he gives a better major premise for Mersenne's argument, and goes on to argue that our clear and distinct perception of God's nature does provide us with knowledge that God's nature is possible. This provides direct textual evidence that "God is possible" is also o-necessary.

about this. I have not claimed that Descartes' modal metaphysics should be understood as invoking alternative worlds in order to explain necessity and possibility, and furthermore such a view would, I think, have a much harder time getting off the ground than the biessentialist view does.[17] Having said that, the ascription of possibility to a proposition does seem, conceptually, to involve a kind of *alternativeness*, as Saint-Germier (2018, p. 4809) puts it, which suggests that if ϕ is possible, then there is (in some sense of "there is") a way for ϕ to have been true. While this needn't be explained by the existence of a possible world, it does seem to permit us to consider and, as far as possible, reason about what would or might have been the case had ϕ been true. This seems to be the case even if we take facts about modality to be reducible to essences, as I have it here. With this in mind, I'll sketch an alternative in which God is not a deceiver, and suggest that it should be ruled out by Descartes' lights.

Suppose, counter-i-possibly, that God had created the world so that different eternal truths were true, and that some of the actual eternal truths were false. If God is o-possibly a deceiver, then He o-could have created the world to have thinking things in it, and furthermore could have given those things the same faculty of conceivability that we enjoy. In this scenario, it could be that a thinker, in employing their intellectual faculties perfectly soundly (i.e. carefully distinguishing their ideas, refraining from judgment until their intellect clearly and distinctly perceives…) may come to error. That is, they may conceive that some proposition is necessary (say, that "2+2=4") which, given the creation of the world, is false. This conception may be perfectly clear and distinct, understood by their intellect, and they may go on to employ their will and come to believe it, perhaps in a way which is "wholly free" (AT 7:58, CSM2 p. 40). In so doing, they land in error through no fault of their own. This is precisely the scenario Descartes seeks to rule out in the Fourth Meditation defense of the truth rule. Since such a scenario is o-possible when God is

[17] Though there is a remark in *The World* (AT 11:47, CSM1 p. 97) involving some discussion of what might have happened if God had created many worlds, this is not clearly about possible worlds in the sense usually meant today, and this passage is not invoked in the argument I'll give here.

o-possibly a deceiver, we, and Descartes, should conclude that God is o-necessarily not one.[18]

This argument is presaged by Spinoza in his *Metaphysical Thoughts*, Part II, Chapter IX (Spinoza 1985, pp. 332–333), where he writes "if God had created things in another way, he would at the same time have constituted our nature so that we would understand things just as they had been created by God." Some care must be exercised in determining how exactly to read this passage, but it seems plausible to me that Spinoza is deducing a consequence of Descartes' views, as Spinoza understands them. In any case, the argument Spinoza gives is similar to that which I give here, regardless of his intentions in composing the salient passage.

The fact that conceivability is coextensive with i-possibility does not mean that Descartes is committed to the claim that conceivability is coextensive with o-possibility. In fact, he seems to commit himself to the view that there are o-possible, inconceivable propositions. This is evidenced in the text of the 2 May 1644 letter to Mesland. God has created our minds, and our faculty of conceivability, to match what he chose to make possible (what is i-possible) not what he could have made possible, but didn't (what is o-possible).[19] A more general upshot for the interaction between this reading of CD and Descartes' epistemic project is that for the purposes of vouchsafing certain knowledge, i-necessity is necessity enough. Put in slightly different terms, the mere o-possibility of some i-impossible ϕ doesn't have an adverse impact on our epistemic lives. We're not epistemically worse off for not being able to conceive propositions which are i-impossible but o-possible.

[18] This argument is meant as a supplement to the point that God's non-deception is o-necessary in virtue of His nature, but is natural enough, and so closely related to the point of the Fourth meditation, that it seemed to me worth concluding.

[19] I assume that the i-possibility of ϕ entails the o-possibility of ϕ, and thus ϕ is o-possibly i-possible only if ϕ is o-possible. (This also involves the assumption that both modalities obey the **S5** properties – in particular, the assumption that possible possibility entails possibility, for both i- and o-possibility.) So it does not introduce a confusion to read 'what God could have made possible' in this passage from the correspondence with Mesland as 'what is o-possible,' rather than as 'what is o-possibly i-possible.'

A fuller discussion of a bimodal approach to questions concerning conceivability and possibility in Descartes is available in Saint-Germier (2018), which provides a compelling account of the upshots of a bimodal account for Descartes' modal epistemology. The treatment of the issue here is to provide some bimodal reasons for admitting, against Frankfurt, the existence of some o-necessary propositions (against those delivered by the essentialism itself), and to give some indication of how conceivability fits into the picture I endorse.

6. Conclusion

In this paper I have developed a *biessentialist* interpretation of Descartes' creation doctrine. I have presented the basic view, providing textual evidence where possible, and giving, hopefully plausible, conjectures where not. In so doing, I have sought to improve on the bimodal picture by arguing that the pair of modalities used to resolve the contradiction in the creation doctrine can be explained in essentialist terms, and furthermore that this is a good fit with Descartes' overall metaphysical commitments. The account has room for improvement, and I cannot claim that it is watertight, but it seems to me plausible and, if it, or a variation on it, works, then it will allow us to see the creation doctrine as a natural part of Descartes' broader metaphysical commitments.

Acknowledgements

I have had very helpful discussions on early versions of this work after talks at UConn, CUNY, University of Melbourne, Ruhr University Bochum, and the Czech Academy of Sciences. Thanks are especially due to a Jc Beall, Florent Dumont, Grace Paterson, Ellie Ripley, Lionel Shapiro, Stewart Shapiro, and Pierre Saint Germier.

References

Alanen, Lilli. 1985. "Descartes, Duns Scotus and Ockham on Omnipotence and Possibility." *Franciscan Studies*, 45: 157–88.

Alanen, Lilli. 1991. Descartes, Conceivability and Logical Modality. In *Thought Experiments in Science and Philosophy*, edited by T. Horowitz and G.J. Massey, 65–84. Lanham, Maryland: Rowman and Littlefield.

Bennett, Jonathan. 1994. "Descartes's Theory of Modality." *The Philosophical Review*, 103(4): 639–77.

Cleve, James van. 1994. "Descartes and the Destruction of the Eternal Truths." *Ration*, (7): 58–62.

Correia, Fabrice. 2012. "On the Reduction of Necessity to Essence." *Philosophy and Phenomenological Research*, 84(3): 639–53.

Cronin, T.J. 1960. "Eternal Truths in the Thought of Descartes and of His Adversary." *Journal of the History of Ideas*, 21(4): 553–59.

Curley, Edwin M. 1984. "Descartes on the Creation of the Eternal Truths." *The Philosophical Review*. 93(4): 569–97.

Descartes, René. 1985. *The Philosophical Writings of Descartes*, vol. 1, translated by John Cottingham, Robert Stoothoff, and Dugald Murdoch. Cambridge: Cambridge University Press.

Descartes, René. 1984. *The Philosophical Writings of Descartes*, vol. 2, translated by John Cottingham, Robert Stoothoff, and Dugald Murdoch. Cambridge: Cambridge University Press.

Descartes, René. 1991. *The Philosophical Writings of Descartes*, vol. 3, translated by John Cottingham, Robert Stoothoff, Dugald Murdoch, and Anthony Kenny. Cambridge: Cambridge University Press.

De Rosa, Raffaella. 2011. "Rethinking the Ontology of Cartesian Essences." *British Journal for the History of Philosophy*, 19(4): 605–22.

Douglas, Alexander. 2020. "Spinoza's Metaphysical Thoughts and the Theological Implications of Cartesian Metaphysics." Preprint. https://www.academia.edu/2094846/Spinoza_s_Metaphysical_Thoughts_and_the_Theological_Implications_of_Cartesian_Metaphysics. Accessed 07.2020.

Embry, Brian. 2017. "Francisco Su´arez on Eternal Truths, Eternal Essences, and Extrinsic Being." *Ergo*, 4(1): 557–78.

Fine, Kit. 1994. "*Essence and Modality." Philosophical Perspectives, 8, Logic and Language*, 1–16.

Frankfurt, Harry. 1977. "Descartes on the Creation of the Eternal Truths." *The Philosophical Review*, 86(1): 36–57.

Gaukroger, Stephen. 1989. *Cartesian Logic*. Oxford: Oxford University Press.

Geach, Peter. 1973. "Omnipotence." *Philosophy*, 48: 7–20.

Gueroult, Martial. 1953. *Descartes selon l'ordre de raisons*. Paris: Vrin.

Hale, Bob. 2013. *Necessary Beings: An Essay on Ontology, Modality, and the Relations Between Them*. Oxford: Oxford University Press.

Hale, Bob. 2018. "The Basis of Necessity and Possibility." *Royal Institute of Philosophy Supplement*, 82: 109–38.

Ishiguro, Hide. 1986. The Status of Necessity and Impossibility in Descartes. In *Essays on Descartes' Meditations,* edited by Amelie Oksenberg Rorty, 459–71. Berkeley: University of California Press.

Kaufman, Dan. 2002. "Descartes' Creation Doctrine and Modality." *Australasian Journal of Philosophy*, 80(1): 24–41.

Keefe, Rosanna and Jessica Leech. 2018. Essentialism and Logical Consequence. In *Being Necessary: Themes of Ontology and Modality from the Work of Bob Hale*, pp. 60–76, Oxford: Oxford University Press.

Kenny, Anthony. 1968. *Descartes: A Study of His Philosophy*. South Bend, Indiana: St. Augustine's Press.

McFetridge, Ian G. 1990. Descartes on Modality. In *Logical Necessity and Other Essays*, edited by John Haldane and Roger Scruton, 155–212. *Aristotelian Society Series*, vol. 11.

Pessin, Andrew. 2010. "Divine Simplicity and the Eternal Truths: Descartes and the Scholastics." *Philosophia*, 38(1): 69–105.

Plantinga, Alvin. 1980. *Does God Have a Nature?* Milwaukee, Wisconsin: Marquette University Press.

Priest, Graham. 2016. *Towards Non-Being*, 2nd ed. Oxford: Oxford University Press.

Routley, Richard. 2018. *Exploring Meioning's Jungle and Beyond The Sylvan Jungle*, vol. 1, edited by Maureen Eckert. Cham: Springer.

Rozemond, Marleen. 2008. Descartes's Ontology of the Eternal Truths. In *Contemporary Perspectives on Early Modern Philosophy: Essays in Honour of Vere Chappell*, edited by Paul Hoffman, David Owen, and Gideon Yaffee, 41–63. Peterborough, Ontario: Broadview Press.

Saint-Germier, Pierre. 2018. "Conceivability, Inconceivability, and Cartesian Modal Epistemology." *Synthese*, 195: 4785–4816.

Secada, Jorge. 2000. *Cartesian Metaphysics: The Scholastic Origins of Modern Philosophy*. Cambridge: Cambridge University Press.

Shalkowski, Scott A. 2004. "Logic and Absolute Necessity." *The Journal of Philosophy*, 101(2): 55-82.

Spinoza, Benedict de. 1985. *The Collected Works of Spinoza*, vol. 1, translated by Edwin Curley. Princeton, New Jersey: Princeton University Press.

Spinoza, Benedict de. 1994. *A Spinoza Reader – The Ethics and Other Works: Benedict de Spinoza*, translated by Edwin Curley. Princeton, New Jersey: Princeton University Press.

Suárez, Francisco. 1983. *On the Essence of Finite Being as Such, On the Existence of that Essence and Their Distinction*, translated by Norman J. Wells. Milwaukee, Wisconsin: Marquette University Press.

Organon F 32 (4) 2025: 457–485 ISSN 2585-7150 (online)
https://doi.org/10.31577/orgf.2025.32404

RESEARCH ARTICLE

Artifacts' Intention-Essentialism and The Production of Negative and Queer Objects

Adrián Solís*

Received: 8 December 2024 / Revised: 15 June 2025 / Accepted: 27 June 2025

Abstract: Artifacts are mind-dependent objects; they exist by virtue of the intentional activity of their makers, at least according to what philosophers usually defend. Some philosophers consider artifacts to be causally and essentially dependent on the intentions of their original makers—a position that we will call *Artifacts' Intention-Essentialism*. Although this position is very attractive, it suffers from various problems. This paper focuses on a discussion of two of those problems: the possibility of creating negative, and queer objects. The first criticism was originally raised by Evnine (2016) against Thomasson's view, but our aim is to defend this criticism for all *Artifacts' Intention-Essentialism* views, including Evnine's. The second problem is a new issue regarding how easy it would be to create queer kinds of artifacts if we accept *Artifacts' Intention-Essentialism*.

Keywords: Artifacts, intention-dependent objects; negative actions; negative objects; negative properties; queer objects.

* Universitat de Barcelona
https://orcid.org/0009-0006-4407-5698
LOGOS Research Group in Analytic Philosophy, Departament de Filosofia, Universitat de Barcelona, Carrer Montalegre, 6-8, 08001, Spain.
solisadriansolis@gmail.com

1. Artifacts' Intention-Essentialism

Ontological questions regarding the nature of artifacts are frequently pondered by contemporary metaphysicians. It is customary to define artifacts as human-made objects, intentionally crafted to fulfil certain purposes or functions (Hilpinen, 1992).[1]

In Section 1, we will present the most extensive view concerning artifacts, which we will call *Artifacts' Intention-Essentialism,* being focused on the Thomasson (2007), Baker (2007), and Evnine (2016) works, and we will introduce a criticism raised by Evnine (2016) regarding some variations of this view. Next, in Section 2, we will explore the problem when extended to Evnine's own view and also to Baker's views varieties of *Artifacts' Intention-Essentialism.* Finally, in Section 3, we will introduce a new criticism of *Artifacts' Intention-Essentialism* in a similar vein to that previously discussed.

We will discuss an important view regarding the essence of artifacts, which argues that the intention of the original makers is an indispensable component of the artifact's nature;[2] we will refer to this as *Artifacts' Intention-Essentialism* (AIE). Various authors subscribe to AIE's view by considering the agent's intention as essential to artifacts. Despite internal differences among AIE views, all of them consider that the intentions play an indispensable role in the nature of the artifacts. All of them consider that without intentional production there are no artifacts because they are intention-dependent objects.

To explain these AIE views we will focus on two important phenomena: the prototype production (which is the first token of a new kind), and the ready-mades (which is the production of a new object without any physical modification of their previous matter). We will focus on the work done by Baker (2000, 2004, 2007), Thomasson (2003a, 2003b 2007, 2009, 2014), and

[1] Some authors do not accept a clear-cut between natural objects, as organisms, and artifacts. See for example Koslicki (2018, ch.8). For a criticism to substantiality of artificial kind, see Preston (2023).

[2] See Elder (2007) for a view in which artifacts are causally mind-dependent but they are not constitutive mind-dependent. See Preston (2013) for a used-based account of artifacts.

Evnine (2016, 2019, 2022), who are the most prominent Artifacts' Intention-Essentialism proponents.[3]

In order to introduce this type of view, consider the following from Thomasson:

> The intention to make something of kind K thus must be based on a declarative intention associating that kind with a number of criteria that would constitute success at creating a K, and involving a number of K-relevant features that the inventor intends to impose on the object in order to succeed at producing a K. This fits in naturally with Hilpinen's emphasis [1992, 641] on the fact that genuine artifacts must have a number of intended properties, thus requiring makers to have a structured intention regarding a number of properties they intend to impose on the object (thus not just a bare intention to create "one of these"). Later makers may acquire the concept K through acquaintance with the initial prototype or with later copies, or they may independently arrive at it, but as long as their intention shares the same content as the original, and they intend to impose the same K-relevant features, they may be said to have the required intention to create a K. (2003b, 597).

The artifacts are considered as entities that are causally and constitutively dependent on the maker's intentions (Thomasson 2007, 53; Baker 2007, 11; Evnine 2016, 69). This is because artifacts are the sort of objects that would not exist without agents with propositional attitudes and the appropriate intentions to produce these objects, since nature does not produce those kinds of objects. Additionally, the essence of artifacts are the intentions of their makers, who stipulate the conditions of existence of the kinds of artifacts and decide which of those characteristics individuate the artificial kinds. Those characteristics could be functional, aesthetic, or structural... All artifacts are created by the imposition of an intention to a piece of

[3] There are other important works that could be classified as Artifacts' Intention-Essentialists like Dipert (1993), Levinson (2007), Houkes and Vermaas (2010), Pearce (2016), Paek (2023), Juvshik (2021). However, we do not have space to discuss every proposal. We select Baker, Evnine, and Thomasson proposals to our discussion.

matter, on the contrary, a random production without an intended action is just a by-product (Thomasson 2007, 67; Evnine 2022, 6).

This is why proponents of AIE consider that “artifacts are the imposition of the makers mind onto the matter” (Evnine 2016, 70) because the artifacts are the product of the intended actions of their original makers, where the resulting product reflects the intentions of their makers (Baker 2007, 54). AIE views are different in their details, but all of them share the view that intentions are indispensable for the essence of artifacts. Additionally, there are two important phenomena of AIE views that they share, and which will be important to the criticisms of AIE. The AIE proposals will be challenged by the combination of these two important phenomena.

The first concerns the possibility of ready-mades[4]. This refers to the production of a new artifact without the necessity of modifying the pre-existing material from which it is made. The most iconic example of a ready-made object is the case of a coffee table made from driftwood. Imagine that you find a piece of driftwood in the forest that has an appropriate structure to be used as a coffee table. You proceed to take the piece of driftwood home with you in order to produce a coffee table with the piece of driftwood but without modifying it, simply using the object as a coffee table. For AIE views this is a genuine case of an artifact’s production, in which an agent had the appropriate intentions to create a coffee table, then a new artifact comes into existence (i.e., coffee table), which is numerically different from that from which it is made (i.e., piece of driftwood)[5] (Thomasson 2007; Baker 2007 43–44; Evnine 2022, 3).

The production of ready-made objects is linked to another important phenomenon: the prototype production. What happens when there are no precedent tokens of artifacts of a kind? How do we reproduce the established

[4] Sometimes, it is also called a “found-object,” but that refers to the same sort of object. For criticisms to ready-made objects, see Zimmerman (2002) or Effingham (2010).

[5] There are other iconic examples of ready-made, such as *The Fountain* of Duchamp, where a numerically new artifact comes into existence from a urinal without any physical modification of the previous artifact (Evnine 2013; 2016, 133–39). However, among AIE views, not all of them accept artworks as artifacts because they do not have any proper function associated (Baker 2007).

conditions of existence of the original makers? According to Thomasson, to produce an artifact is to have the largely successful intention of producing an artifact of the relevant kind *K*. Thus, for an agent to produce an artifact, the agent must know what the K-relevant features of the kind *K* are that she intends to produce, and then have the intention of producing them (2003b, 592–596). As we have said, these K-relevant features may be very different. If there are no previous tokens of a new kind, AIE views have some differences between them. For Thomasson (2007, 60–61), in order to produce a new kind of artifact, we need to fulfil the associated conditions of existence and a successful application of the concept towards a piece of matter. However, according to Baker (2007, 56) and Evnine (2016, 122–124), there is another important requirement: the resulting product of the intentional activity must be capable of performing the associated function of the new kind of artifact. Then, according to AIE, in order to create the first token of a new artifact, all we need is the intentional application of a concept about the conditions of existence (Thomasson 2007), and the resulting product is capable of performing the associated function to the new kind (Baker 2007; Evnine 2016). We also have to take into consideration that there are no restrictions to producing a prototype with a ready-made object.

The main point of AIE's proposals is that, regardless of the fact that an artifact *A* could perform a very different range of functions due to its physical characteristics, the artifact *A* essentially belongs to the kind *K* because it was intentionally produced by the subject *S* to be a *K*. For instance, an ice axe is essentially an ice axe because the maker intended to produce an object of the kind *ice axe*, independently of the fact that the same ice axe may also perform other functions not primarily associated with the kind *ice axe*. For instance, when, on 21 August 1940, Ramon Mercader (a Catalan communist spy) used an ice axe to kill Trotsky, the ice axe was performing the function of a weapon but could also perform others, like *bottle-opener*, etc. In this case, Ramon Mercader did not create a weapon, he simply used an existing object, an ice axe, as another weapon because of the physical characteristics of the ice axe. What makes an artifact primarily belong to a certain artificial kind is having been an intended product of this kind.

1.1. The Negative-Production Problem

Our aim is to discuss a particular criticism in the debate. The criticism was raised by Evnine (2016) to Thomasson's proposal. The criticism is as follows:

> Chairs [in Thomasson's view] exist when people work with certain intentions on appropriate matter not, as I have it, because those people are exercising some creative power to bring chairs into existence thereby, but simply because the conditions contained in the concept obtain. If we define a concept *not-thair* in such a way that it applies just when a person ignores some potential chair-matter and does nothing to it, then we can say that not-thairs exist (and have always existed) in the same way, under conditions that clearly are not genuinely creative in the first-order sense. (2016, 117)

Evnine considers Thomasson's view of artifacts to be overly ontologically permissive. Considering that Thomasson defends an ontological minimalism view of objects, where all that is needed for an object to exist is to apply adequately the concept to a piece of matter (or something else), Evnine criticises Thomasson for portraying the creation of a new object not as the product of a creative act, but merely as a satisfactory application of the concept (2016, 117). Thus, if all that is required to produce objects is to apply the concepts, and if the concept is correctly applied, not-thairs would exist in the world in the same sense that chairs exist. However, although Evnine argues that objects like not-thairs do not exist because they lack a creative act (2016, 117), nevertheless, in Evnine's proposal, artifacts also depend causally and constitutively on the makers' intentions. He contends that the criticism does not apply to him because the production of these things as not-thairs is not a creative action (2016, 117). Evnine mentions the distinction between creative and non-creative processes only to differentiate his proposal from Thomasson's, but he does not explain it at all. In any case, Evnine does not seem satisfied with the existence of artifacts such as not-thairs.

However, Evnine's approach to solving the problem is far from clear. Koslicki (2018, 223) points out that Evnine does not provide good reasons to dismiss the criticism. Firstly, Koslicki notices that Evnine leaves open the essence of mental states like intentions, which is essential for distinguishing between creative and non-creative intentions. Secondly, he cannot simply appeal to the non-existence of not-thairs to explain how an agent cannot succeed in having a creative intention with not-thairs. Thirdly, he also cannot appeal to the faultiness of the concept of not-thair because he needs to explain what makes a concept faulty. These considerations lead Koslicki to conclude that Evnine cannot explain why the not-thair criticism is not a problem for his proposal regarding the nature of artifacts (2018, 223). Koslicki does not develop this criticism, and our aim will be to discuss in a deeper way this sort of criticism of Evnine's view.

In the next section, we will show that the criticism that Evnine makes of Thomasson, is a problem for his own proposal and other proposals similar to him[6] (e.g., Baker's view). To do this, we will first introduce some ways in which Evnine could consider that the not-thair is a non-creative case, and then we will show why these reasons are not good reasons to block the criticism. Cases like not-thair will be an important problem for these Artifacts' Intention-Essentialism proposals.[7,8]

2. Bringing Negative Objects into Existence?

The examples above pose a problem for Artifacts' Intention-Essentialism by introducing controversial objects, such as not-thair, which Evnine discusses. If someone aims to defend an AIE view, they must explain how to

[6] Evnine does not take in consideration a possible reply that Thomasson is able to give based in her late proposal of artifacts in (2014) where she introduces external restriction to the artifacts' creation based on the public norms of artifact kinds. However, our aim is to show that Evnine and Baker's views have that problem, independently of the Thomasson responses.

[7] See Kornblith (2007), Koslicki (2018, 2023), Eaton (2020) or Solís (2024) for other criticisms to Artifacts' Intention-Essentialism views.

[8] There are other proposals that considers the intentions make an important role on the essence of the artifacts; we cannot discuss all these versions, we will focus just on Thomasson (2003), Baker (2007), and Evnine (2016) views of the Artifacts' Intention-Essentialism.

address these cases, and this is what Evnine attempts to justify. Previously, we mentioned how Koslicki (2018, p. 223) considers that Evnine cannot resolve the not-thair cases. However, we contend that Koslicki's criticism of Evnine's response is not completely accurate. This is because she does not attempt to elucidate why Evnine is saying that there is no creative making in the not-thair case. Our current task will be to present two different possible responses that Evnine could make, based on what his actual proposal says about how new artifacts come into existence. Nevertheless, we will explain why these two possible responses from Evnine are not good enough. We sympathise with Koslicki's goal in criticising Evnine, but we differ in the means.

Let us to introduce two different ways in which Evnine could justify that the not-thair cases do not involve creative intentions:

(a) **No-Action Strategy:** This case involves not doing something (specifically, ignoring the creation of a chair); it is an absence of an action. To create something, you need to work on the matter, either by making physical changes, or by using ready-mades.

(b) **No-Kind Strategy:** This case involves only a concept, not a new kind of object. To create a new kind of artifact it is necessary that the maker has in mind the concept of the new kind *K*', imposing it onto some matter, and that the product is something that is able to perform the associated function that the maker establishes in the concept of *K'*.

(a) and (b) are two different ways in which Evnine could address the criticism based on his proposal. The first involves his consideration that the creative act—the individual essence of the artifact (2016, p. 109)—comprises two different components: the intention, and the labor. The latter is the work done by the maker on the matter, guided by the intention, to bring the artifact into existence. The labor is an action performed by the maker on the matter (2016, p. 70), and the not-thair case could be interpreted as an absence of labor. (b) is based on his assumptions about prototype production, which involves an agent formulating a new kind of object with certain existence and persistence conditions and an associated function. The first new exemplar of the kind must be able to perform its associated function (2016, pp. 124-5). Thus, it is not sufficient to have a concept about something. What is needed is something more robust: a new artificial kind.

In order to ease the discussion, we will refer to objects like not-thair as negative objects, and we will define it as:

> **Negative Object (NO):** an object which its essence consists in a negative property, such as *not being suitable to sit*, *not being suitable to write*, *not being red*, etc.[9]

In subsequence sections, we will first show why the possibility of negative objects is a problem for Evnine's view, and second, why it is a problem for other views that defend the intentions that figure in the essence of artifacts, as Baker's (2007) view. The main point of the criticism is to show that if we accept AIE views on the nature of artifacts and the phenomena of ready-mades and prototype production, we will arrive at non-desirable creations such as negative and queer artifacts.

2.1. Refusing No-Action Strategy: The Case of Negative Actions

We have said that Evnine, and other AIE proponents, could use different ways to block the criticisms of negative objects like not-thair. Let us focus first on the No-Action Strategy. It states that just ignoring some potential chair-matter to be a chair is not an action; there is no labor involved where the maker *works* on the matter. As there is an absence of action, there is no creation. In Evnine's proposal, artifacts like chairs are the product of the imposition of a maker's mind onto the matter (2016, p. 89). The

[9] A negative object is capable of having positive properties, such as *being red, being spherical*, etc. However, its essence is a negative property. Consider that essence is a real definition (in Finean's (1994) sense because not every necessary negative property is essential to an object, e.g. numbers are colourless but this does not mean that numbers are negative objects, because this property is not the essence of the numbers). The real essence of a negative object would be an entity which consist in the negative property such as *not being able to sit* (plus other contingent positive properties). Others define negative objects as consisting of solely negative properties (see Hommen, 2018, 401). It is not our case.

essence of artifacts, such as a chair, is the original creative act that the maker imposes on the matter. This act of creation is divided into two components:[10] the intention of the maker, and the labor through which the maker imposed his intention. The No-Action Strategy suggests that in the not-thair scenario, there is no labor performed by the agent, because the agent is just "*ignoring* some potential chair-matter, and *he does nothing* with it" (2016, 117; my emphasis). Evnine could argue in this scenario, that there is no action involved because there is no labor of imposing an intention into the matter; the agent is simply ignoring a piece of matter and doing nothing with it. We should now consider whether this strategy is a suitable way to argue that the scenario does not involve a creative intention, as Evnine intends to, or if there are good reasons to assert that ignoring can count as a creative action.

Let me introduce a scenario involving different sort of actions that resemble the case of not-thair introduced by Evnine. It is the case of negative actions: On December 1st, 1955, Rosa Louise McCauley (Rosa Parks) returned from work, boarded the usual bus, and chose a seat in the designated area for black people. During the journey, several white people entered the bus, and the bus driver, wanting to ensure the comfort of white passengers, instructed black individuals, including Rosa Parks, to vacate their seats. However, Rosa Parks refused to comply. She intentionally decided not to stand up and remained seated. Subsequently, the bus driver called the police, and the officers arrested Rosa Parks because she refused to stand up.

In the Rosa Parks scenario, how many actions did she take on the bus? She made physical movements such as getting on the bus and sitting on a bus seat. However, what happened when she refused to stand up from her bus seat? Can we say that she performed an action of *not doing something*? Or can we say that she was not engaging in an action; that her refusal was just a mere occurrence?

Some authors consider that cases such as that of Rosa Parks reveal genuine actions that have not received much attention: negative actions, which are "exercises of agency that seem to consist primarily in an agent

[10] He notices that this distinction may just be a conceptual or ontological one, but he does not pretend to discuss it (2016, 70).

not doing a certain thing" (Payton, 2021, 2). Those who defend the existence of negative actions such as omissions or refusals, consider these actions to exist alongside positive actions. We will call them Negative Actions Realists (NAR). They aim to defend the existence of negative actions because they are causally efficacious (Bernstein, 2015, 212–215; Payton, 2018, pp. 95–100; Palmer, 2020, 744). For example, Rosa Parks' refusal to stand up caused her arrest and trial (if she had not been physically able to stand up, she would not have been arrested and judged). NAR also seeks to rationalise the agent's behaviour when performing the omission (Payton, 2018, 89). For instance, what explains Rosa Parks' refusal to stand up when the bus driver notified her is that she intends *not to stand up*. NAR argues that negative actions have a structure very similar to uncontroversial entities, such as positive actions (Payton, 2018; 2021) or material objects (Palmer, 2020). Just as a lump of clay constitutes a statue, Palmer argues that a positive action constitutes the negative action (2020, 741–743).

For Negative Actions Realists, negative actions are considered a subset of actions and have the same level of reality as positive actions. Both positive and negative actions exist in the world and are determined by the intentions—whether positive or negative—that agents have. By regarding negative actions as real phenomena in the world, we can more accurately formulate our criticism of Artifacts' Intention-Essentialism.

Let us consider that negative actions are real actions, following the reasoning of Negative Actions Realists. How does this affect Evnine's resistance to the not-thair case, considering the explanation in (a) about why the not-thair case does not count as being creative? (a) suggests that in the not-thair case, the agent is merely omitting something, and that she is not doing anything by omitting to do something. However, if we accept NAR, when the agent omits to do something, it is not an absence of an action; it is an action, just like a positive action. Evnine cannot then argue that in this scenario there is no labor because there is no action. In fact, there is an action, which is a negative action to *not doing something*. He cannot then argue that this scenario is not creative because there are not actions. The agent is doing something (i.e., not doing something else).

2.1.1. Considering Possible Replies

It is possible that Evnine may not be persuaded about the parallels between what the agent does in the scenario of not-thair, and the cases of negative actions such as Rosa Parks not standing up. There are different ways in which Evnine can refuse to accept that his view implies the existence of not-thair: (i) to ignore is not necessarily intentional, or (ii) negative actions are not creative.

In (i) he could say that ignoring is not necessarily an intentional activity because there are cases in which we say that a person is ignoring something, but merely because the person does not know about it. Allow me to share some comments about this.

First, the scenario introduced by Evnine has a particular goal: to serve as a criticism of Thomasson's view on the essence of artifacts. The main point is that for Thomasson, the essence of an artifact is the intentional application of a concept to a particular quantity of matter. Given this, Evnine considers that we can create concepts, the result of which is an application-concept that is a very rare entity such as a not-x. The main point is that if the Thomasson view is committed to an intentional application of the concept by the agent involved in the production of the artifact, then the main action of the stipulated concept cannot be unknown to the agent involved. Thus, 'ignoring' is an intentional activity in the not-thair scenario if it is to be a criticism of Thomasson's view. We must be aware that Thomasson is defending a view in which the essence of an artifact is the intentional activity of the original maker that applies concepts adequately, and Evnine is criticising Thomasson's view because he considers that she is overly liberal about the sufficiency of the intentional activity. By that reason, if the scenario that Evnine presents there are not intentional activity by the agent involved then it would not be a counterexample to Thomasson. If we read "ignoring" as unintentional activity, then Thomasson can refute the criticism by saying that it is not a case of artifact creation because unintended products of human activity are not artifacts.

Secondly, Evnine does not pay much attention to explaining what "ignoring" is in the scenario that he presents. Thus, we must understand "ignore" in the ordinary and standard sense, which is an intentional activity. For example, consider the definition of "to ignore" provided by the Oxford

English Dictionary: "To refuse to acknowledge (a person or thing); to disregard intentionally" (2023). Evnine does not propose a revisionary sense of "ignoring". It seems, then, that we should assume that what Evnine is referring to as "ignoring" is its ordinary meaning.

Regarding (ii), he could accept the existence of negative actions, and that in the not-thair scenario there is a negative action involved (i.e., to ignore the creation of a chair), even though this action is not a creative one. He could argue that creative acts necessarily involve an intention and a labor, but it is not the case that whenever you have an intention and a labor, this entails a creative act; negative actions cannot serve as creative acts. Let me say why (ii) it is not a good defence either.

Firstly, it would be an odd strategy because when Evnine discuss the creative acts he does not define what it is to be creative and what it is to not be creative; he just says that "the action of making whereby mind is imposed on matter is both *necessary* and *sufficient* for determining the individual essence of an artifact" (2019, 203; my emphasis). The intention and the labor are sufficient to determine the creative act of the agent who tries to produce an artifact. Evnine cannot appeal by saying that negative actions are different to positive actions because only the latter are creative; he does not have independent reasons for this conclusion.

Secondly, he cannot just say that negative actions are not creative because there is an absence of a definition of creativity; he cannot justify which actions are creative or not if he does not explain what creativity is. When Evnine is referring to creative acts, he is considering it with a *generative power*, in which something new comes into existence. Nonetheless, saying that creative acts are generative, do not exclude why a negative action are not generative. However, let us to consider what happens if we take into consideration the different existent definitions of *creativity* in the specific literature:[11]

Let us start by considering the Valuable View, in which creativity is associated with the notion of *genius*, as when Kant defines the artistic genius as someone that produces things which are original and exemplary (2000, 182–197). To be creative is to be valuable in some sense (Klausen,

[11] We will not enter into a full discussion of all the proposals of creativity; we will just present them. For more detail, see Paul and Stokes (2023).

2010). Evnine does not seem to be working with this view because, in his view, to say that an artifact is not valuable does not mean that the object is not a product of a creative act. For example, he distinguishes between valuable and ordinary artifacts. In the case of valuable objects, he argues that objects that result from artisanal production are valuable because they are the unique product of a singular act of creation. In the case of ordinary objects (non-valuable), he includes mass-production artifacts like thumbtacks, where many thumbtacks can result from a singular creative act (2019, 199). There are ontological differences between the two because artisanal artifacts have individual essences, whereas mass-production artifacts have only collective essence. However, both involve creative acts in which something numerically new emerged (2016, 100; 2019, 209).

Consider, then, the Surprise View (Boden, 2010) in which creativity occurs when the maker intends that his product is to be surprising: not that the product per se is surprising, but the maker in the process of production tends it to be surprising when he creates something new. Regarding Evnine's view that there is no importance about the surprise in any sense, cases of mass-production involve creative acts, and they do not seem focused on anything surprising; the worker in the factory is just making repetitive bodily movements to produce pre-established things (2016, 97–103).

Moving onto other views, consider the Originality View (Kronfeldner, 2009), in which creativity involves some original process of making. Evnine does not subscribe to this view because if we consider the creation of a table, it is not necessary that the agent produce the artifact through an original process of making in order for his action being creative. For the act of the creation of a table, it is sufficient that the agent has in mind the concept of table-kind, having the intention of produce it by working adequately on the matter. Originality is not necessary in the process of making.

If we consider the Spontaneity View (Kronfeldner, 2009), in which creativity involves unintentional activity, this sort of view cannot be what Evnine has in mind when he talks about creative acts because the most important element in his view is that the artifacts are the imposition of a maker's mind onto matter through an intentional activity.

Nonetheless, there is a more attractive view of creativity which has many supporters who fit with Evnine's view: the Agency View (Carruthers, 2006; Gaut, 2018; Kieran, 2014; Stokes, 2014). According to this view, the creative process must involve the intentional activity of the agent. For other views, it is possible that something counts as creative without the intentional intervention of an agent, and these authors consider the intentional activity of agents as indispensable. This view fits with Evnine's explanation about creative acts, which always involve an intention, and an action guided by this intention, to create this sort of object. This is because, for Evnine, artifacts are ideal objects because they are the imposition of a maker's mind onto matter (2016, 69)

Although Evnine could argue that he accepts the existence of negative actions in the not-thair scenario, but also that they are not creative, this is not a sound strategy because he has never explained what creativity is for him. If we consider the available literature regarding the debate of creativity, the only view that fits with his proposal is the Agency View, which considers that creativity is the intentional activity of the agent, and thus, negative actions, in virtue of being intentional, would count as genuinely creative. In this case, Evnine's reply is not a good one.

Thus, a No-Action Strategy does not seem to be the best way to argue that the not-thair case does not apply to Evnine's proposal. Let us consider whether a No-Kind Strategy is a better option.

2.2. Refusing No-Kind Strategy: Turning Not-Chair into Existence

Let us to focus on the second option, the No-Kind Strategy. We can now argue that this strategy is also flawed and that Evnine's view entails the possibility of the production of negative objects. We defined a negative object as:

> **Negative Object (NO):** an object in which its essence consists in a negative property, as *not being suitable for sitting*, *not being suitable for writing*, *not being red*, etc.

With this definition of what a negative object is, we will show why Evnine's view entails the existence of these objects, and why these objects are undesirable in the ontology.

First, remember what the No-Kind Strategy says: the not-thair case is just a case about the stipulation of a concept which applies in certain circumstances. The scenario described does not challenge Evnine's proposal because it does not involve the introduction of a new kind of artifact. The not-thair is not a novel prototype with an associated function. To count as a real creation, we need a kind of object with an associated function that at least, in the first instance of the new kind, is able to perform the function (Evnine 2016: 123–124) (albeit if in the later instances it cannot perform the function. If the maker intentionally tries to create this kind of object, the resulting object will be an exemplification of the kind; a broken *K* is still a *K*) (Evnine, 2016, 127). Thus, the not-thair is not a problem for Evnine's AIE view because the application of concepts is not sufficient for the artifact's production; we need something additional.

AIE views have a way to resist that the not-thair case can be a problem to their proposals, but this is only because not-thair is defined as being a concept that applies under certain circumstances. However, we will consider another case that is very similar and is inspired by the not-thair example,[12] but one which implies a kind of artifact, not just a concept. Assuming Evnine's view about the prototype, we will show that nothing prevents this kind of controversial creation.

The relationship between kinds, prototypes, and functions in Evnine's view is important. To solve the problem of shared functions among different kinds of artifacts (2016, 123–124), an artifact's kind is not determined by what it is actually able to do, but by what it is supposed to do (2016, 220). Remember the ice axe, which could be used as a weapon but in fact was not a weapon. This artifact belongs primarily to the kind *ice axe*, with the associated function of 'allowing climbers to fix themselves onto the ice', etc. Evnine considers that an artifact is what it is supposed to do, independently of whether the object can perform the function associated with its kind. A knife is still a knife if it is broken, but it is produced with the intention to be a knife (2016, 120). However, considering the case of a prototype—the introduction of a new kind of artifact—what we need is that the original maker defines the associated function that the new object will have, and

[12] We may consider the cases as analogous, but this could be questioned. In any case, the newly introduced case is very similar and is inspired by the not-thair.

assuming a more objective criterion, the resulting product must be able to perform the function associated with its kind (2016, 126). Different kinds of artifact have different functions associated with them, determined by their original authors and their subsequent makers; to bring into existence new tokens they must intentionally produce them to perform the associated function of the kind.

With this information, we will consider that Evnine's view is overly liberal with his considerations about the nature of artifacts, and his conception opens the door to the production of controversial entities. We will consider a similar case to a not-thair scenario, but which implies a new kind of artifact. Consider the new kind of artifact:

> *Not-chair* $=_{\text{df}}$ it is a kind K that has the associated function of *not being suitable for sitting.*

This not-chair is a new kind of artifact that has the associated, or proper, function of not being suitable to sit on. We have previously defined a negative object as a kind of object whose essence is a negative property. The not-chair would be an exemplification of a negative object. Now, consider that until this moment, nobody had the intention to produce a not-chair. However, imagine that at this moment an agent A now has the intention of producing an object of the kind *not-chair.* It will be a novel prototype; the first instance of a new kind of artifact. As an artisan, A knows the kind of object that she desires to produce. She selects some raw materials and she works on the matter, guided by the intention to produce a token of the kind *not-chair*, which will be an artifact which will be impossible to sit on. If, at the end of her work on the matter, the resulting entity is something on which it will be impossible to sit (in virtue of the matter selected, or the arrangement of the material parts, etc.), then it seems that she has been successful in the creation of the negative object *not-chair.* This is a problem for Evnine's view because the process of creation seems to be the imposition of the maker's mind onto the matter, and subsequently, the imposition of the kind not-chair into some matter. The demand for a novel prototype of being capable of performing the associated function is thus achieved.[13]

[13] Note that in the definition of not-thair as a concept, the production of the *not-thair* was necessarily by ready-mades; the agent just omitted to produce a chair.

Therefore, Evnine's view entails the possibility of producing negative objects such as not-chairs. This illustrates that Evnine's view attributes a significant amount of power to the maker in order to produce a new artifact. There is no specific limitation to creativity; negative objects could also be brought into existence if someone desires it and acts accordingly.

2.2.1. Considering A Possible Reply

Nonetheless, it is possible that Evnine does not see this as a substantive criticism of his proposal. He could reply by saying that the ontology contains more objects than we usually acknowledge, and that makers have the capacity to produce artifacts of a very wide variety. He can "bite the bullet" and say: there are chairs, and not-chairs—or there could be in our ontology—all of which are intentional productions of makers with associated functions, and that is what artifacts are.

Let me make some comments about this sort of defence. (i) it would go against the grain of their realistic budgets[14], and it would entail a revisionary ontology; he is just posing the existence of negative artifacts before showing that his view entails them; he would be inflating the ontology unnecessarily. (ii) it needs the existence of negative properties that do not exist (Armstrong, 1978, 1989; Strawson, 1974; Vallée, 2004), or they do not have the same degree of existence as positive properties (Zangwill, 2011).

Let us consider (i). If Evnine's view accepts that there are, or could be, artifacts which are negative objects because they could be an intentional production with an associated function, then he has just given an answer to the problem that contradicts his meta-ontological framework to construe

Now, in the definition of not-chair that we propose, the not-chair production could be modelling the matter in some way (by performing positive actions) or by ready-mades or found-objects by referring to an already created artifact or a natural object, which would have available a physical structure to *not being suitable for sitting*, and referring intentionally to this object by saying that it will be a not-chair (as in the case where a coffee table is created from a piece of driftwood).

[14] Evnine could adopt a permissivists account of objects, but in fact he is not a permissivist, because he denies it (2016, 195–201). He tries to defend a view between permissivism and nihilism, proximate to commonsense account (2016, 14).

a position between permissivism (2016, 14) and nihilism which highly counterintuitive implications. We do not ordinarily accept that there are such artifacts as not-chairs just because someone wants to make something on which it is not suitable to sit. If Evnine intends to accept that negative objects exist as well as ordinary objects, he would be inflating the ontology unnecessarily. There will be in the world a plethora of objects that we do not know exist. We could say that he would be multiplying entities without necessity; with the existence of negative objects, he would be violating Ockham's Razor.

(ii), if we consider the existence of negative objects (e.g., not-chairs) which have a negative property (not being suitable to be sat on) as their essence, then we have to assume that there are negative properties in the world and that they have, at least, the same degree of existence as positive properties that define the essence of other uncontroversial objects, such as chairs. However, if we take into consideration the literature about the existence of negative properties, the standard view considers that they do not exist, which is defended by Strawson (1974), Armstrong (1978, 1989), and Vallée (2004). Moreover, for those who accept the existence of negative properties in the world, these properties do not have the same causal and determinative power that positive properties have (Zangwill, 2011). An object can have negative properties; for example, a blue chair can have the negative property of *not being orange*. However, negative properties are always predicated on a positive property; they do not have determinative power for individuated entities in the world (Zangwill, 2011). Therefore, as negative properties lack causal and determinative power, they do not seem good candidates for being the essence of artifacts like not-chairs. In Evnine's view, by biting the bullet with negative objects, we would want to say that not-chairs exist as well as chairs. However, as said negative properties do not exist in the same degree of reality as positive properties, negative objects, in the best case, do not exist in the same way as positive objects. Perhaps Evnine could solve this problem by explaining why negative properties exist in the same way that positive properties do. However, he does not explain this in any sense; the onus to explain is his. Currently, his view contains no explanation about the existence of negative properties. Thus, following the standard view in which these properties do not exist (Strawson

(1974); Armstrong (1978; 1989), Vallée (2004))–or at least do not exist with the same determinative power (Zangwill (2011))–negative objects like not-chair would be a problem because they would entail the existence of negative properties.

Evnine can bite the bullet with the existence of not-chairs, but if he did that, it will be a strategy that will contradicts his ontological project with no independent justification, because they would be inflating the ontology without necessity, with counter-intuitive implications for ordinary experience, and without further support for the existence of negative properties associated with these kinds of objects.

2.3. A Coda: Expanding The Problem Baker's Artifacts' Intention-Essentialism

In the previous sections, we discussed the Negative-Production Problem directed at the Evnine view. We will now explain why this problem is not exclusive to Evnine and Thomasson's views but is also a problem for another important variety of Artifacts' Intention-Essentialism.

Consider now Baker's view. She also defends the Artifacts' Intention-Essentialism, but unlike Thomasson, she considers that all artifacts have proper functions in virtue of the kind that they belong to; this is what makes an entity an artifact. Her view is usually seen as focusing only on technical artifacts: artifacts with proper functions (Baker, 2007, 49–51). As we have said, she accepts the possibility of ready-mades—those new artifacts that are created without modifying the pre-existing material from which they were made (2007, 43–44)—and also that for prototypes, what we need is that the resulting product reflects the intentions of their makers (2007, 54). Giving that her proposal is more restrictive than Thomasson's view because it adds an additional requirement to the intentional activity of the maker (i.e., the resulting object is able to perform its proper function), then we could say that she would deny that the not-thair scenario was a problem for her. This is because, using the No-Kind Strategy, she could say that the scenario only involves a new concept, but this concept does not have a proper function. Thus, the resulting product does not fulfil the conditions of a new kind of artifact.

Nevertheless, if she uses this sort of defence, she will run into the same problem that we previously presented with Evnine's view. Since her proposal gives too much power to the makers for the production of new artifacts, and if we consider the new artifact's kind as not-chair, she will be forced to accept their existence. Remember the not-chair case: the not-chair has the proper function of *not being suitable for sitting.* If there is a maker with the intention of producing a not-chair through a not-chair favourable circumstance—the maker chooses the appropriate matter, works on it intentionally, and the resulting product satisfies the maker's intentions because the resulting product is an artifact which is *not suitable for sitting* due to the materials used, the structure of the materials… then there would be not-chairs in our ontology, which is an undesirable consequence for the reasons we previously mentioned.

Therefore, scenarios such as not-thair*,* presented by Evnine, or a similar one that we presented on the *not-chair*, reveal a problem for the Artifacts' Intention-Essentialism because their proposals entail the possibility of the existence of these negative objects because the existence of these kinds of objects contradicts their ontological projects that pretends to preserve the commonsense view (Evnine 2016, 14; Baker 2007, 5–7, 25–26, 98)[15] (in the case of Evnine and Baker's views).[16]

3. Further Queer Objects Coming into Existence

We have seen how AIE views on the nature of artifacts are problematic because it would be possible to produce artifacts that would belong to an undesirable category of negative object, just as long as the original makers had the intention to produce such things. There is nothing in their proposals

[15] In the case of Thomasson's view this might not be seen as a problem in virtue of their meta-ontological permissivists considerations of easy ontology (2009). However, Evnine made this sort of criticism in virtue of the consideration that these objects would be so much controversial (2016, 117). In any case, our criticism is directed to Evnine (2016) and Baker (2007) AIE's views.

[16] Notice that in the original criticisms, in Evnine (2016) to Thomasson's view, (and in Koslicki (2018) to Evnine's view) involves ready-mades' cases, but the discussion of the case that we are offering ready-mades are not necessary but also involves these cases.

to stop this sort of production. Moreover, they do not seem to be in a good position to bite the bullet, for all the counterintuitive implications that this might have. We will now introduce another criticism to AIE that is similar to those previously discussed.

According to AIE, for the production of new kinds of objects it is sufficient to define (i) the condition of existence, (ii) the function associated with it, and (iii) the prototype that exemplifies (i) and (ii). Given that, consider the following kinds of artifacts:

> *Chairtable* $=_{\mathrm{df}}$ it is an intentional production that results in a material thing with some physical properties which can perform the function of *being able to be sat on* (what a chair is) **and** *to put things on* (what a table is).
>
> *ChairOR* $=_{\mathrm{df}}$ it is an intentional production that results in a material thing with certain physical properties that performs the function of a chair **or** the function of x (e.g., a computer), and the resulting material object performs the function of a chair.

These kinds of artifacts show that we can define a kind of artifact as having randomly complex properties, such as conjunctive or disjunctive properties. Thus, if a maker has these kinds in mind in order to produce an artifact, and the product of her activity is a material artifact that can perform the randomly associated complex function, then this kind of artifact would come into existence. For example, in the case of a *chairtable*, the kind involves a conjunction, where the resulting object must perform the two functions involved in virtue of the truth-conditions of the conjunction. A maker can then create an object that is qualitatively the same as a chair and can perform the functions of sitting on and *putting things on* because the chair, through its physical properties, serves both functions.

In the case of *chairOR*, the kind involved is a disjunction in which the truth-conditions are sufficient for one of the disjuncts to be true. Then, if the maker has in mind to produce an artifact that performs the function of a chair or a computer, and she produces an indistinguishable material object as a chair, she will be producing a *chairOR*, not a chair. This is because the product of her intentional activity has the function of a chair or a computer. By the truth-condition of disjunctions, it is true if at least one of the disjuncts is true. If the product of her activity performs one of the functions of the disjunct (i.e., chair-function), then the artifact is a *chairOR*, not a chair, although they are qualitatively indistinguishable.

However, are there actually artifacts in the world like *chairtables* or *chairOR*? Ordinarily, we would say that there are no such kinds of objects. AIE views open up the possibility of inflating the ontology just by subjective considerations. All that we need to create new artifacts in the world is to invent a kind, produce it intentionally, and have the resulting object perform the random function attributed to it. AIE views are too liberal regarding artifact production. Perhaps the conditions they establish for producing artifacts are necessary, given intentional production, but they obviously are not sufficient, as seen in the different examples that we gave. We need a more restrictive, realistic, and objective proposal on what artifacts are.

Considering the cases presented as *chairtable* or *chairOR*, they show that by assuming Artifacts' Intention-Essentialism, it would be possible to produce objects that are qualitatively indistinguishable from an ordinary chair (colour, shape, structure, etc.), but that they would be essentially different from an ordinary chair simply because the content of the original maker's intentions was different. There would be nothing different *in* the object if we were to consider a *chair* and a *chairOR,* but they would be essentially different artifacts; they would belong to different artificial kinds. Note that the cases presented are not different ways of talking about artifacts like chairs. What they show is that just by subjective considerations of introducing new queer kinds of objects, it would be possible to create objects that are qualitatively identical to ordinary kinds of artifacts, but they would be essentially different just by the subjective considerations due to the content of the intentions of their makers. Ontological questions about artifacts could depend on queer subjective considerations.

Again, the problem with these kinds of objects is similar to the previous criticism of the production of negative objects. If AIE views bite the bullet on the creation of queer objects, as previously mentioned, then they would be inflating the ontology of objects without further justification. Their assumption is completely *ad hoc*, and highly counterintuitive since we do not recognise these kinds of objects in our ordinary experience or taxonomy. By accepting the AIE view, there may be a possible identification problem: an agent faced with a piece of matter that seems qualitatively indistinguishable from a chair, could be completely mistaken about what kind of object it is, because if the original maker of that object had the intention to produce a *chairOR* that is qualitatively indistinguishable from a chair, there is no way for other agents to identify what kind of artifact the resulting object is. With the AIE view, we would have an ontology in which, for qualitatively indistinguishable objects which are internationally produced, one is a chair, and the other is a chair-wherever. Again, AIE's view would be violating Ockham's Razor by multiplying entities without necessity. *Chairtable* and *chairOR* would be qualitatively indistinguishable from ordinary chairs, then it seems that we do not need those kinds of objects in our ontology; it is enough to have chairs.

4. Implications and Future Directions

The Artifacts' Intention-Essentialism is the most extensive and important view about the essence of artifacts. Although it has multiple advantages, it suffers from various criticisms. In this work, we have seen two new criticisms for this sort of view. The first criticism is based on a previously existing one that was directed to an Artifacts' Intention-Essentialism view–Thomassons' view–and following Koslicki's scepticism about whether this criticism also applies to Evnine's view, we have extended the criticism in a more sophisticated way to Evnine and other Artifacts' Intention-Essentialism points of view. Our criticism shows how Artifacts' Intention-Essentialism entails the existence of negative objects, and how all attempts to block this criticism are unsatisfactory. The second criticism is new and follows the line of reasoning of the previous one, which is that Artifacts' Intention-Essentialism could entail the existence of queer kinds of objects such as kinds with conjunctions or disjunctions. Based on the considerations for the prototype productions of Artifacts' Intention-Essentialism, these kinds of queer objects could exist if an agent has the intention to produce them. Both criticisms show how easily it is to create undesirable kinds of objects, such as negative and queer ones would be produced if we accept Artifacts' Intention-Essentialism; our ontology would be inflated unnecessarily just by subjective considerations. We should, therefore, reject Artifacts' Intention-

Essentialism and move towards other proposals that do not give that much power to individual agents in the production of artifacts.

We consider this could suggest two possible alternatives: (a) abandoning the framework of AIE and building a proposal in which artifacts do not essentially depend on their makers' intentions or (b) reducing the ontological implications of individual makers by accepting a limited framework of the AIE. Option (a) is suggested by Koslicki and Massin (2025), they focus on giving greater importance to the capacities of artifacts. Option (b) could be an interesting way to explore the essence of artifacts, in which they do not depend only on individual intentions but shared intentions of social communities, where artifacts' existence may not be akin to other objects, such as organisms.

However, both options must face significant challenges. These two options propose revisionary approaches to how we ordinarily conceive the existence of artifacts. Option (a) prioritises *functionality* or *capacities* over makers' intentions, which could be seen as contrary to the usual talk about artifacts. Option (b) reduces the role of individual makers' intentions and gives greater status to the shared intentions of communities, perhaps through *recognition* or *acceptance.*

At the same time, it remains unclear whether options (a) and (b) would entail the same, or even greater, criticisms for Artifacts' Intention-Essentialism. Nonetheless, AIE has different significant problems that suggest the move to explore alternative ontological proposals that reduce the ontological power of the individual makers. Those could be represented by focusing the essence on the *capacities* or *functions* of the artifacts or focusing the essence on the *recognition* or *acceptance* of a social community, in any case further investigation would be necessary to evaluate these possible ontological routes.[17]

Acknowledgements

I am indebted to different people, who help me in different ways in the production process. This work has been informed by the conversations held, at different

[17] I must thank to reviewers for prompting me to suggest the ontological implications of the criticisms that we have introduced in this work.

stages of the project, with Marta Campdelacreu. I am also indebted to the reviewers, whose comments and suggestions have made me improve some aspects of the article. Finally, my most sincere debt is to my family. This work has been possible because Júlia Estrada has listened to it, discussed it and read it countless times, and the philosophical depth of her criticisms have made me improve it. Diana Casabella Estrada and Sara Casabella Estrada have always been in the middle listening and discussing my crazy philosophical ideas. In addition, I am deeply grateful to Diana for her linguistic corrections that have undoubtedly helped in the clarification of my ideas.

Funding

Work of this publication has been supported by the following grants: CEX2021-001169-M funded by MICIU/AEI/10.13039/501100011033; PID2023-150569NB-I00 funded by Ministerio de Ciencia, Innovación y Universidades de España; PID2023-150066NB-I00 funded by Ministerio de Ciencia, Innovación y Universidades de España.

References

Armstrong, David. 1978. *A Theory of Universals*, vol. 2. Cambridge: Cambridge University Press.

Armstrong, David. 1989. *Universals: An Opinionated Introduction*. Boulder, Colorado: Westview Press.

Baker, Lynne Rudder. 2000. *Persons and Bodies: A Constitution View.* Cambridge: Cambridge University Press.

Baker, Lynne Rudder. 2004. The Ontology of Artifacts. In *On Human Persons*, edited by K. Petrus, 23–39. Frankfurt am Main: Ontos Verlag.

Baker, Lynne Rudder. 2007. *The Metaphysics of Everyday Life: An Essay in Practical Realism*. Cambridge and New York: Cambridge University Press.

Bernstein, Sara. 2015. “The Metaphysics of Omissions.” *Philosophy Compass*, 10(3): 208–218. DOI: https://doi.org/10.1111/phc3.12206

Boden, Margaret A. 2010. *Creativity and Art: Three Roads to Surprise*. Oxford and New York: Oxford University Press.

Carruthers, Peter. 2006. *The Architecture of the Mind: Massive Modularity and the Flexibility of Thought*. Oxford: Clarendon Press.

Clarke, Randolph. 2014. *Omissions: Agency, Metaphysics, and Responsibility*. Oxford and New York: Oxford University Press.

Dipert, Randall. 1993. *Artifacts, Art Works, and Agency.* Philadelphia: Temple University Press.

Eaton, A. W. 2020 Artifacts and Their Functions. In *The Oxford Handbook of History and Material Culture*, edited by Anne Carter and Ivan Gaskell, 36–53. Oxford: Oxford University Press.

Effingham, Nick. 2010 "The Metaphysics of Groups." *Philosophical Studies*, 149: 251–67. DOI: https://doi.org/10.1007/s11098-009-9335-4

Elder, Crawford. 2007. On the Place of Artifacts in Ontology. In *Creations of the Mind: Theories of Artifacts and Their* Representation, edited by E. Margolis and S. Laurence, 33–51. Oxford University Press: New York.

Evnine, Simon. 2013. "Ready-Mades: Ontology and Aesthetics." *British Journal of Aesthetics*, 53(4): 407–23. DOI: https://doi.org/10.1093/aesthj/ayt033

Evnine, Simon. 2016. *Making Objects and Events: A Hylomorphic Theory of Artifacts, Actions, and Organisms*. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780198779674.001.0001

Evnine, Simon. 2019. "Mass production." In *The Nature of Ordinary Objects*, edited by Javier Cumpa and Bill Brewer, 198–222. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781316612897.010

Evnine, Simon. 2022. "The Historicity of Artifacts: Use and Counter-Use." *Metaphysics*, 5(1), 1–13. DOI: https://doi.org/10.5334/met.74

Fine, Kit. 1994. "Essence and Modality." *Philosophical Perspectives*, 8: 1–16. DOI: https://doi.org/10.2307/2214160

Gaut, Berys. 2018. The Value of Creativity. In *Creativity and Philosophy*, edited by B. Gaut and M. Kieran, 124–29. New York: Routledge.

Hilpinen, Risto. 1992. "Artifacts and Works of Art." *Theoria*, 58(1): 58–82. DOI: https://doi.org/10.1111/j.1755-2567.1992.tb01155.x

Hommen, David. 2018. "Negative Properties–Negative Objects?" *Acta Analytica*, 33(4): 395–412. DOI: 10.1007/s12136-018-0341-z

Houkes, Wybo and Vermaas, Peter E. 2010. *Technical Functions: On the Use and Design of Artifacts*. Dordrecht: Springer.

Juvshik, Tim. 2021. "Artifacts and mind-dependence". *Synthese* 199(3-4): 9313–36. https://doi.org/10.1007/s11229-021-03204-6

Kieran, Matthew. 2014. Creativity as a Virtue of Character. In *The Philosophy of Creativity: New Essays*, edited by S. E. Paul and S. B. Kaufman, 125–44. Oxford/New York: Oxford University Press.

Klausen, Søren H. 2010. "The Notion of Creativity Revisited: A Philosophical Perspective on Creativity Research." *Creativity Research Journal*, 22(4): 347–360.

Kornblith, Hilary. 2007. How to Refer to Artifacts. In *Creations of the Mind: Theories of Artifacts and Their Representation*, edited by Eric Margolis and Stephen Laurence, 138–149. Oxford and New York: Oxford University Press.

Koslicki, Kathrin. 2018. *Form, Matter and Substance.* Oxford: Oxford University Press.

Koslicki, Kathrin. 2021. "The Threat of Thinking Things into Existence." In *Commonsense Metaphysics: Essays in Honor of Lynne Rudder Baker* edited by L. R. G. Oliveira and K. J. Corcoran, 113–36. New York: Routledge.

Koslicki, Kathrin. 2023. Artifacts and the Limits of Agentive Authority. In *Philosophers in Depth: Thomasson on Ontology* edited by M. Garcia-Godinez, 209–41. London and New York: Palgrave/Macmillan.

Koslicki, Kathrin and Massin, Oliver. 2025. Artifact-Functions: A Capacity Based Approach. In *Special Objects: Social, Fictional, Modal, and Non-Existent*, edited by María J. García-Encinas and Fernando Martínez-Manrique, 31–51. Cham: Springer.

Kronfeldner, Maria E. 2009. "Creativity Naturalized." *The Philosophical Quarterly*, 59(237): 577–92

Levinson, Jerrold. 2007. Artworks as Artifacts. In *Creations of the Mind: Theories of Artifacts and Their Representation*, edited by Eric Margolis and Stephen Laurence, 74–82. Oxford: Oxford University Press.

Paek, Chaeyoung. 2023. "Making Things Collectively." *Metaphysics*, 6(1): 1–12. DOI: https://doi.org/10.5334/met.91

Palmer, David. 2020. "Omissions: The Constitution View Defended." *Erkenntnis*, 85(3): 739–56.

Paul, Elliot S. and Stokes, Dustin. 2023. Creativity. *Stanford Encyclopedia of Philosophy*, Edward N. Zalta & Uri Nodelman (eds.). https://plato.stanford.edu/archives/spr2023/entries/creativity/

Payton, Jonathan D. 2018. "How to Identify Negative Actions with Positive Events." *Australasian Journal of Philosophy*, 96(1), 87–101. DOI: https://doi.org/10.1080/00048402.2017.1297843

Payton, Jonathan D. 2021. *Negative Actions: Events, Absences, and the Metaphysics of Agency.* Cambridge and New York: Cambridge University Press.

Pearce, David. 2016. "Collective Intentionality and the Social Status of Artifactual Kinds." *Design Science*, 2, e3. DOI: https://doi.org/10.1017/dsj.2016.3

Preston, Beth. 2013. *A Philosophy of Material Culture: Action, Function, and Mind.* New York: Routledge.

Preston, Beth. 2022. "The Artifact Problem: A Category and Its Vicissitudes." *Metaphysics*, 5(1): 51–65. DOI: https://doi.org/10.5334/met.86

Solís, Adrián. 2024. "La Artefactificación: Un nuevo problema para el Esencialismo Intencional de los Artefactos." *Revista Portuguesa de Filosofia*, 80(3): 943–78. DOI: 10.17990/RPF/2024_80_3_0943

Stokes, Dustin R. 2014. The Role of Imagination in Creativity. In *The Philosophy of Creativity: New Essays* edited by E.S. Paul and S.B. Kaufman, 157–84. Oxford and New York: Oxford University Press.

Strawson, Peter F. 1974. *Subject and Predicate in Logic and Grammar.* London: Methuen and Co.

Thomasson, Amie. 2003a. "Foundations for a Social Ontology." *Protosociology*, 18–19: *Understanding the Social II: Philosophy of Sociality*, 269–90. DOI: https://doi.org/10.5840/protosociology200318/199

Thomasson, Amie. 2003b. "Realism and Human Kinds." *Philosophy and Phenomenological Research*, 67(3): 580–609. DOI: https://doi.org/10.1111/j.1933-1592.2003.tb00309.x

Thomasson, Amie. 2007. Artifacts and Human Concepts. In *Creations of the Mind: Theories of Artifacts and Their Representation* edited by Eric Margolis and Stephen Laurence 52–73. Oxford and New York: Oxford University Press. DOI: 10.1093/oso/9780199250981.003.0004

Thomasson, Amie. 2009. Social Entities. In *Routledge Companion to Metaphysics*, edited by Robin LePoidevin, Peter Simons, Andrew McGonigal, and Ross Cameron, 545–54. Routledge: London. DOI:10.4324/9780203879306-57

Thomasson, Amie. 2014. Public Artifacts, Intentions, and Norms. In *Artefact Kinds: Ontology and the Human-Made*, edited by Pieter Vermaas et al., 45–62. DOI:10.1007/978-3-319-00801-1_4

Vallée, Richard. 2004. "What Is Wrong with Negative Properties." *Manuscrito*, 27(2): 361–82.

Zangwill, Nick 2011. "Negative Properties." *Noûs*, 45(3): 528–56. DOI: https://doi.org/10.1111/j.1468-0068.2010.00776.x

Zimmerman, Dean W. 2002. "The Constitution of Persons by Bodies: A Critique of Lynne Rudder Baker's Theory of Material Constitution." *Philosophical Topics*, 30: 295–338. DOI: https://doi.org/10.5840/philtopics200230111

Organon F 32 (4) 2025: 486–490 ISSN 2585-7150 (online)
https://doi.org/10.31577/orgf.2025.32405

RESEARCH ARTICLE

Is Moral Knowledge Necessary for Moral Worth?

Yuanfan Huang*

Received: 11 May 2025 / Accepted: 8 June 2025

Abstract: The article examines the necessity of moral knowledge for moral worth, focusing on Neil Sinhababu's (2024) arguments. Sliwa (2015) and Cunningham (2021) contend that moral worth requires moral knowledge. In contrast, Sinhababu (2024) challenges this view using Gettier cases, arguing that justified true belief, even without knowledge, can still confer moral worth. This article argues that Sinhababu's Gettier cases do not convincingly demonstrate that moral knowledge is unnecessary for moral worth.

Keywords: Moral knowledge; moral testimony; Gettier case; moral worth.

Regarding the question of whether moral knowledge is a necessary condition for moral worth, Sliwa (2015) and Cunningham (2021) contend that morally worthy actions inherently require moral knowledge. Sliwa posits that a morally right action possesses moral worth only if it is motivated by both a concern for doing what is right and knowledge that it is the right thing to do. Cunningham further argues that morally worthy actions must be motivated by an awareness of how to respond to the reasons that justify the action, which is informed by propositional knowledge of the specific

* Shanghai Jiao Tong University
https://orcid.org/0000-0001-9144-2063
✎ School of Humanities, Department of Philosophy, Shanghai Jiao Tong University, 800 Dongchuan RD. Minhang District, Shanghai, China
✉ huangyuanfan@sjtu.edu.cn

normative reason at hand. In contrast, recently, Neil Sinhababu (2024) suggests that Gettier cases—scenarios in which justified true belief fails to qualify as knowledge—indicate that moral knowledge is not necessary for moral worth.

In the following discussion, I will demonstrate that Sinhababu's Gettier case do not convincingly serve as counterexample to the claim that moral knowledge is necessary for moral worth.

Sinhababu introduces a Gettier case called "Texting the Rabbi" to question whether moral knowledge is necessary for moral worth. In this scenario, Ava and Beth face a moral dilemma involving William, who angrily demands the return of weapons he loaned to them. Both women seek guidance from their rabbi through text messages. Ava receives an authentic response from the rabbi, advising her not to return the weapon, which gives her a true belief grounded in moral knowledge. In contrast, Beth's message is intercepted by a thief, who provides her with a random reply by flipping a coin, leading her to believe that it is right not to return the weapon. Although Beth holds a justified true belief, she does not possess moral knowledge.

Sinhababu argues that Ava's and Beth's actions of not returning the weapon have equal moral worth despite one has moral knowledge while the other lacks. Both faced the same situation, sought advice, and acted on it in the same way. The difference in their knowledge—Ava's belief is based on genuine moral testimony, while Beth's is based on a random response—does not affect the moral worth of their actions. Beth's justified true belief, though not knowledge, still leads to a morally right action. Sinhababu concludes that moral worth does not require moral knowledge. The Gettier case that matter for knowledge are not the same as those that matter for moral worth. Actions motivated by justified true belief, even in Gettier cases, can have moral worth. In other words, one can acquire justified true belief regarding moral issues by luck and still possess moral worth.

Let us delve deeper into Sinhababu's cases by providing more details. In the first scenario, we assume that both Ava and Beth accept the Rabbi's words without further reflection or inquiry. If this is the case, they will embrace his guidance unconditionally, fully deferring their moral dilemmas to him. First, it is crucial to recognize that true moral saints are exceedingly rare in reality. Even those who are generally virtuous can make misguided moral choices. There exists a genuine possibility that the Rabbi may act unethically at times or arrive at flawed conclusions concerning the moral challenges faced by Ava and Beth.[1] By choosing to relinquish their significant moral decisions to someone who is likewise fallible, they risk merely relying on luck should the Rabbi happen to provide sound advice.

One might argue that it is logically possible for the Rabbi to consistently make correct moral judgments and offer sound moral guidance to those who seek it. However, when individuals appeal to God or an external "foundational truth," they are, in Sartre's view, evading the responsibility of creating their own meaning in life. By deferring to a higher authority, they avoid acknowledging that they alone are responsible for their actions. Ava and Beth's reliance on the Rabbi mirrors this, as they forfeit their moral autonomy by accepting his guidance without reflection. In doing so, they shift their moral responsibility onto others, thereby diminishing their own moral worth. We can imagine a similar scenario in the future with an all-knowing moral AI capable of solving complex ethical dilemmas. If Ava and Beth were to always follow the AI's suggestions when uncertain about moral decisions, we would intuitively question their moral integrity in this case as well.[2]

Now, let's examine the second scenario, which is more plausible in everyday life. It is more likely that Ava and Beth would not simply accept the Rabbi's "yes or no" advice without further inquiry; instead, they would seek to understand the reasons behind it. When Beth receives a message from

[1] It is entirely possible for even a moral saint to make poor decisions. Firstly, proponents of situationism challenge the idea of global character traits—qualities that consistently manifest across different situations. For instance, a person may display courage on the battlefield but fail to show the same bravery in a courtroom. Secondly, many complex moral dilemmas may not have straightforward solutions. Consider the case of an individual who is uncertain about whether to use AI to assist with an assignment. If he opts to seek guidance from the Rabbi, there is no guarantee that the Rabbi will be able to address this issue effectively, especially if he is unfamiliar with how large language model (LLM) AI operates.

[2] Compared to non-moral testimony, suppose Jake receives testimony about scientific knowledge *p* from Joe, a scientist who discovered the fact *p*. It is clear that the credit should be given to Joe. Similarly, it seems that the moral worth should be attributed to the individual who makes the moral decision.

the fake Rabbi stating, “You should not return the weapon,” it would be unusual for the conversation to end there. Typically, a genuine Rabbi would provide justifications for such guidance. Therefore, in a typical situation, if the fake Rabbi fails to offer further explanation, Beth would likely ask for additional clarification. With these characterizations, there are two possibilities. In the first scenario, if the fake Rabbi is merely making decisions by flipping a coin, Beth will eventually realize that her message has been intercepted. In that case, she would stop trusting the fake Rabbi’s moral testimony. In the second scenario, if the fake Rabbi wants to continue the interaction and is skilled at giving moral advice—perhaps through extensive reading of moral philosophy—Beth may not realize that the Rabbi is fake. In this situation, the fake Rabbi could be considered as knowledgeable as the real Rabbi when it comes to moral knowing-that; since if he weren’t, he would be easily exposed as an imposter.[3] However, even though the fake Rabbi lacks virtuous character, his possession of extensive moral propositional knowledge might lead one to question why Beth’s trust in his moral advice wouldn’t be justified, given the knowledge he holds.

So far, I have argued that in one scenario, if Ava and Beth accept pure moral testimony without further reflection, neither of them can justifiably be attributed with moral worth. In another scenario, if Ava and Beth require an understanding of the reasons behind the moral testimony, it is likely that Beth would recognize the message as coming from a fake Rabbi, since a fake Rabbi is typically not competent to provide sound moral guidance. However, we must also consider the logical possibility that the fake Rabbi possesses knowledge equivalent to that of a genuine Rabbi in terms of propositional moral knowledge; in such a case, his moral testimony could be considered warranted. Given these considerations, three possibilities arise: either Beth does not deserve moral worth, or she would reject the fake Rabbi’s testimony, or she is justified in believing the fake Rabbi’s claims. In any of these scenarios, Beth’s situation cannot support the claim that the Gettier case entails that moral knowledge is unnecessary for moral worth.

[3] Clearly, the fake Rabbi only possesses “knowing-that” regarding moral issues; therefore, it is possible that he might fail to act according to this knowledge, which requires “knowing-how.”

Funding

This research is sponsored by the National Foundation of Social Science of China (No. 23BZX071).

References

Cunningham, Joe. 2021. “Moral Worth and Knowing How to Respond to Reasons.” *Philosophy and Phenomenological Research*, 105(2): 385–405. https://doi.org/10.1111/phpr.12825

Sinhababu, Neil. 2024. “Moral Worth in Gettier Cases.” *Journal of Ethics and Social Philosophy*, 29(1): 151–58. https://doi.org/10.26556/jesp.v29i1.3129

Sliwa, Paulina. 2015. “Moral Worth and Moral Knowledge.” *Philosophy and Phenomenological Research*, 93(2): 393–418. https://doi.org/10.1111/phpr.12195

Organon F 32 (4) 2025: 491–500
https://doi.org/10.31577/orgf.2025.32406

ISSN 2585-7150 (online)

BOOK REVIEW

Samuele Iaquinto & Giuliano Torrengo: *Fragmenting Reality: An Essay on Passage, Causality, and Time Travel* London: Bloomsbury, 2022, x+208 pages

Giacomo Andreoletti*

1. Introduction: Flow Fragmentalism

This book offers a comprehensive overview and a compelling defense of what the authors label *Flow Fragmentalism*, a particular version of the theory of time known as fragmentalism (cf. Fine 2005, 2006). Stepping back a little, we can divide the various philosophical theories on the nature of time into two broad families: tensed theories of time (or A-theories) and tenseless theories of time (B-theories). Tensed theories take tense to be a fundamental aspect of reality. For instance, an A-theorist would typically say that reality is composed, among other things, of irreducible A-properties, such as the objective *pastness* of Caesar crossing the Rubicon or the objective *presentness* of me typing this sentence. Tensed theories are usually said to be *dynamic*, as they portray reality as changing and passing in a robust way. For instance, now that you are reading this sentence, the objective presentness of me writing *that* sentence is gone and no longer part of reality.

On the other hand, tenseless theories of time hold that alleged tensed aspects of reality can and should be reduced to more fundamental tenseless formulations. Typically, a tenseless theorist would explain away the pastness of Caesar crossing the Rubicon by saying that, on a fundamental level, what we have is not the event of Caesar crossing the Rubicon possessing the property of being past. Rather, it is simply that the event of Caesar crossing the Rubicon is *earlier than* my current temporal perspective, here in 2024, as I am writing this. For a

* University of Salzburg
https://orcid.org/0000-0001-5632-4213
Department of Philosophy, University of Salzburg, Franziskanergasse 1, 5020 Salzburg, Austria
giacomo.andreoletti@protonmail.com

tenseless theorist, the fundamental earlier-than relation among times does most of the explanatory work in accounting for the nature of time. As such, tenseless theories are considered *static* theories of time. To appreciate why, it suffices to note that if event A is earlier than event B, this is always the case and is not subject to change.

Flow Fragmentalism, the view defended in this book by Samuele Iaquinto and Giuliano Torrengo, stands in the first camp. It is a tensed theory of reality. However, it is a *non-standard* form of tense realism. To appreciate the peculiar (and intriguing) character of Flow Fragmentalism and its non- standard aspect, it is useful to contrast it with standard tensed theories. Standard forms of tensed theories include *Presentism* (the view that only what is present exists), the *Growing Block Theory* (the view that what exists is either past or present), and the *Moving Spotlight Theory* (the idea that nothing comes into existence or goes out of existence, yet the fabric of time is "illuminated" by a continuously moving objective present).

Flow Fragmentalism, on the other hand, is a non-standard form of A-theory. I will use a toy example to illustrate their view. Consider a universe composed only of two times, *t1* and *t2*. Socrates is sitting at *t1* and standing at *t2* . What is part of temporal reality according to Flow Fragmentalism? The fundamental ingredients are *tensed* facts – this makes Flow Fragmentalism a tensed theory of time – such as the fact that Socrates IS sitting and the fact that Socrates IS standing. (I use the present tense in capital letters to highlight that we are "taking tense seriously" and not merely describing a B-theoretic fact such as the obtainment of something at a particular time *t*, since such a B-theoretic fact is not part of reality according to the flow fragmentalist.)

A second ingredient of the view is the idea that reality is unique: there is only one reality. This forces the flow fragmentalist to give up coherence. Under Flow Fragmentalism, reality is not coherent, as it contains contradictory facts. Reality is constituted by the fact that (look at *t1* in the example above) Socrates IS sitting, and hence also by the fact that he IS not standing. But reality is also constituted by the fact that (look at *t2*!) Socrates IS standing. Reality is unique and incoherent. However, coherence reappears when we take the perspective of fragments, i.e., maximally coherent collections of tensed facts. Reality is unique (and incoherent), but it comes in (coherent) pieces: the fragments.

To illustrate the fundamental aspect of Flow Fragmentalism, Iaquinto and Torrengo draw a distinction between *constitution* and *obtainment.* In their view, the former is an absolute notion, whereas the latter is a relative one. Tensed facts *constitute* reality "period" – i.e., in an absolute sense and without relativization to times – while tensed facts always *obtain* within a certain fragment. In the example, the fact that Xanthippe IS standing and the fact that she WILL be sitting obtain within fragment$_1$, whereas the fact that she IS sitting and WAS standing obtain within fragment$_2$. Nonetheless, all those facts constitute reality in an absolute sense.

The third ingredient of Flow Fragmentalism is the claim that no fragment is privileged. The different fragments are pieces of a unique reality, and all fragments are born equal, with the same metaphysical standing (cf. p. 39). As Iaquinto and Torrengo emphasize, Flow Fragmentalism endorses the principle of Neutrality: with respect to what facts constitute reality, no time is privileged (cf. p. 20). In other words, all fragments and all times are metaphysically on a par. This is perhaps the crucial element that makes Flow Fragmentalism a non-standard tensed theory, as it brings it somewhat closer to tenseless theories of time. After all, for a B-theorist as well, no time is privileged, and all times are on a par.

2. Book Outline

Here's the outline of the whole book. The introduction provides the setup: the main assumptions are laid out and made explicit, and the reader is introduced to the debate in which Flow Fragmentalism is a contender – namely, tensed vs. tenseless theories and standard vs. non-standard tensed theories. Chapter 1 provides the foundations of Flow Fragmentalism. The theory is carefully presented, and a formal semantics is provided – this is especially useful, as some of the distinctions in Flow Fragmentalism are subtle and nuanced (e.g., obtainment vs. constitution). The formal language the authors provide makes explicit the implications and commitments of the view. Chapter 1 also offers a "derivation" in seven steps showing how Flow Fragmentalism can account for, and most importantly *explain*, the passage of time.

The subsequent chapters address some of the most common topics in the metaphysics of time and explore how Flow Fragmentalism can address them. Chapter 2 deals with the open future. Here, Iaquinto and Torrengo develop a branching-time version of their Flow Fragmentalism. They also make a distinction between the *ontological* openness of the future – roughly, the future is open because it doesn't yet exist – and *topological* openness – there are many futures, hence the openness. The branching version of Flow Fragmentalism brings together elements from both the ontological and the topological conceptions, and it does so once again with the distinction between obtainment and constitution. In a nutshell, their idea is that the future can be open in an ontological sense from the point of view of obtainment – within a fragment, future events do not exist, even though we have the obtainment of future tensed facts – while from the point of view of constitution, we have topological openness – there are multiple futures.

Chapter 3 discusses cross-temporal causation. Here, Iaquinto and Torrengo address an aspect that is *prima facie* problematic for Flow Fragmentalism. Causation seems to be intrinsically a cross-temporal relation. However, tensed facts are understood as being confined and sealed within their respective fragments. How, then, can they interact in a causal manner across different fragments? To overcome this problem, the book proposes introducing a *pseudo* earlier-than relation. Relations between events, including causal relations, are reduced to internal relations that occur in different fragments. Crucially, internal relations are not existence-entailing, and this allows Flow Fragmentalism to account for causal relations without relying on external cross-temporal relations.

Chapter 4 generalizes Flow Fragmentalism to the case of special relativity. Often, discussions in the metaphysics of time assume Newtonian time and its notion of absolute simultaneity for the sake of simplicity in exposition. Here, Iaquinto and Torrengo make the noteworthy effort to adapt their Flow Fragmentalism to relativity. The final chapter deals with the case of time travel, both in the Ludovician (non-past-changing) case and in the non-Ludovician (past-changing) version. Flow Fragmentalism has little to add to the Ludovician case, but it definitely offers advantages over competitors in the case of past-changing time travel. Past-changing time travel models that involve an eternalist ontology coupled with an objective present (e.g., Hudson & Wasserman 2010, Bernstein 2017, and Effingham 2021) typically suffer from the problem of zombies, i.e., people who exist at a time but never have the "privilege" of being illuminated by the objective present. Here, Iaquinto and Torrengo show how Flow Fragmentalism can accommodate non-Ludovician time travel without incurring the problem of zombies who never were, nor will ever be, present.

As shown, the book does many remarkable things. The main objective of the book, though, is to provide an explanatory account of the passage of time. This is arguably the main strength of Flow Fragmentalism and its non-standard nature. Standard versions of tensed theories, Iaquinto and Torrengo argue, posit the passage of time as a primitive aspect of temporal reality. Consider, for instance, standard dynamic presentism and a frozen version of presentism – i.e., one in which there is only a present time, but it never ceases to be, nor is it replaced by a "new" present. The presentist, in order to avoid the collapse of their theory into frozen presentism, needs to posit as a primitive the fact that there is a "self-propelled" and freeze-avoiding present that constantly makes reality dynamic. Flow Fragmentalism, on the other hand, has the theoretical resources to explain the phenomenon of passage. This is the main advantage of Flow Fragmentalism over standard rivals.

I will not enter here into the details of the account of the passage of time in Flow Fragmentalism, but let us see some of the key ideas. What explains passage in Flow Fragmentalism? As mentioned, Iaquinto and Torrengo want to avoid a picture in which the present is a sort of "unmoved mover", i.e., something that brings dynamicity to reality without itself being moved (or "pulled and pushed", in their terminology) by something else, which in turn would be capable of explaining the movement of the present. Rather, they exploit the non-standard aspects of the theory to explain passage. Given that, in their view, all fragments have the same metaphysical status (given neutrality), they are able to explain the obtainment of a future tensed fact within a fragment by virtue of the obtainment in the future of the corresponding present tensed fact—likewise for past-tensed facts. More precisely, Torrengo and Iaquinto account for passage with *Fragmentalist Flow.*

> **Fragmentalist Flow**. Within fragment$_{x}$, TENSE$_{n}$ ϕ because within fragment$_{x+n}$, ϕ.

The operator TENSE here is a shorthand for the irreducibly tensed metric operators WAS and WILL, while *n* refers to the temporal interval between fragments. Following the example of Socrates sitting at *t1* and standing at *t2*, as an instance of Fragmentalist Flow, we'd have that "within *fragment$_{1}$*, WILL(1) Socrates stands because within *fragment$_{2}$* Socrates stands." Similarly, in the direction of the past, we'd have that "within *fragment$_{2}$*, WAS(1) Socrates sits because within *fragment$_{1}$* Socrates sits." What is crucial in the principle and its instances is the explanatory role of the because operator. Each fragment finds grounds for its tensed facts from what obtains in the corresponding fragments. This allows Iaquinto and Torrengo to say that, from the perspective of each time, "there is a push from the past and a pull from the future," which captures and explains the passage of time (cf. p. 40).

Before moving to some potential objections, let me stress some remarkable aspects of this book. For starters, the chapter on relativity theory, where

Iaquinto and Torrengo make their theory compatible with relativity, is particularly noteworthy. This is not only because of the quality and clarity of the chapter, but also because analytic metaphysics of time does not engage enough (or not always) with our best scientific theories. One would expect metaphysics to be least compatible with our best science, but this is not always the case. Some work in metaphysics does not engage (or does not engage enough) with scientific theories. It is thus particularly noteworthy that the book develops a relativistic version of Flow Fragmentalism.

Another aspect that I find remarkable is how the authors are transparent and precise about their methodology, assumptions, and goals of their argumentation. The introduction is a point of reference in this respect. The assumptions are laid out in a clear manner, and the context of the debate is presented succinctly but clearly. Moreover, Iaquinto and Torrengo lay out the details of the methodology of their scientifically informed metaphysics in great detail.

The third remarkable aspect (though the list could go on) is the rigor and precision with which the arguments in the book are carried out. The arguments are clear and cogent, and it is (relatively) easy to follow them. Moreover, when an argument is controversial, Iaquinto and Torrengo acknowledge it, while also considering what an opponent could say. This adds to the overall argumentative strength of the book.

3. Objections

In this section, I want to raise three potential objections. I think that all of them are related to the following. As we have seen before, Flow Fragmentalism is a non-standard view. Its non-standardness derives from the fact that Flow Fragmentalism is a hybrid view – it takes elements from standard tense realism (fundamental tensed facts, for instance) as well as elements from tenseless views (see the principle of neutrality). In general, hybrid views have the advantage of being able to take the best from two opposing worlds and provide a synthesis that avoids problems which pester the opposing camps. Hybrid views, if you will, embrace the Aristotelian maxim that virtue lies in the mean. On the other hand, a potential problem of hybrid views is that they risk collapsing into one of the two sides. I will argue that their account of passage, in my view, risks collapsing into anaemic passage – "passage" which is nothing more than the earlier-than relation.

The first worry I have is rather small, while the second one may be more substantial. The book talks about the *specialness* and the *egalitarian* intuitions.

The first refers to the idea that the time you are experiencing right now is special or privileged. This is the idea, dear to the A-theorist, that there is a metaphysically privileged time, i.e., the present. On the other hand, the egalitarian intuition is the idea, dear to the B-theorist, that all times are born equal and are metaphysically on a par – there is no privileged time. What I find odd is not that Iaquinto and Torrengo seem to want to do justice to *both* intuitions. After all, Flow Fragmentalism is a hybrid view, and as such, it must locate itself in the middle of the two camps. What I find odd is rather that they treat what they call the "egalitarian intuition" as an *intuition* in the first place, in order to then argue that their view does justice to both intuitions (cf. p. 7). I would argue that there is no such thing as the egalitarian intuition. Most (almost all?) people do not seem to have the intuition that all times are equal, or all equally existing, or all on a par. Typically, one would pre-theoretically think that only the present is special. Considerations about egalitarianism regarding times come from either scientific theories (relativity) or philosophical and B-theoretic considerations. As a consequence, I do not see how it could be a virtue of the theory to capture the egalitarian *intuition.*

My next objection is about the account of passage that Flow Fragmentalism provides. As we saw above, the account is based on the principle of Fragmentalist Flow (see section 2). The account has many merits, but I nonetheless want to raise an objection against it. In short, I will argue that the fragmentalist passage collapses into anaemic passage. Consider this passage from the book (p. 40): "…the flow of time is not given by a self-propelled present, but rather by the fact that from the perspective of each time there are a push from the past and a pull from the future that explain how reality is globally updated."

The main advantage of the view, as I see it, is that it does not take passage as an unexplainable primitive. Flow Fragmentalism *explains* passage in terms of Fragmentalist Flow and its explanatory "because" operator. However, I wonder whether the resulting picture provides a satisfactory explanation of passage.

To illustrate my criticism, consider a finite universe constituted by, say, 10 fragments. Suppose fact A obtains in *Fragment$_5$* and in no other fragment. Given Fragmentalist Flow, we have that within fragment$_4$, WILL$_1$ A because within *fragment$_5$*, A. Likewise, we have that within fragment$_6$, WAS$_1$ A because within *fragment$_5$*, A. These causally explanatory links provide the "pushes and pulls" – this talk of pushes and pulls should be taken as more than a metaphor, given its frequent appeal – providing reality with the dynamic feature of passage. *Fragment$_5$* is pushed from the past and pulled from the future. But consider that the view also includes Neutrality: All fragments are metaphysically on a par, and none of them is privileged. So, *Fragment$_4$* is as well pushed from its past and pulled from its future, *Fragment$_6$* is pushed from its past and pulled from its future, and so on. All fragments are equally pushed and pulled, so they are all "moving" in the same direction and at the "same speed." But if they are all "passing" in the same way and for the same reasons, it seems as though, from a global perspective, reality (the totality of fragments) features no passage at all. Reality is just static.

To illustrate the point, consider an analogy with an actual object moving in space. Suppose the object is composed of 10 parts, and they all move with the same velocity. Absent an absolute frame of reference, we could equally describe the 10 parts as moving or as simply being static, depending on the frame of reference. There is no intrinsic movement in the 10 parts. Similarly, for the fragments that compose a universe, it is conceptually difficult to imagine an absolute background against which the passage of time could be said to obtain.

Lastly, I want to consider Iaquinto and Torrengo's response with respect to the Update Test (cf. p. 43–45). In essence, the test runs as follows: Take a theory and the reality it is supposed to describe. Write down how reality is according to the theory. Wait some time. Write down again how reality is according to the theory. The test is passed if the second description is different from the first one. The test is meant to check whether a theory genuinely captures robust passage. Clearly, a standard B-theory does not pass the test – descriptions of reality will feature statements like "Socrates sits at *t1*, Socrates does not sit at *t2*," and that's not going to change. Standard A-theories, on the other hand, will pass the test: A presentist, for instance, would first describe reality as featuring Socrates sitting, wait some time, and then provide a different description, one featuring Socrates not sitting.

As Iaquinto and Torrengo point out, Flow Fragmentalism does not pass the test. Both in terms of constitution and obtainment, reality is always described in the same way. We would have, always, that the fact that Socrates sits and the fact that Socrates does not sit both constitute reality, and that within *fragment$_1$*, the fact that Socrates sits and WILL not sit both obtain. This description is not going to change over time. Thus, the flow fragmentalist description of reality does not pass the update test. To this, Iaquinto and Torrengo reply that the update test is unfair. But I am not sure their arguments against the fairness of the test are fully convincing. To undermine the legitimacy of the update test, they use the following analogy: A theory about a constantly updating reality does not need to be itself constantly updating, just as a theory about vagueness does not need to be itself vague (cf. p. 45). I am not sure this analogy is fitting. Granted, a theory about microscopic reality does not need to be itself microscopic, but I'm not sure this has much bearing on the distinction between static and dynamic theories of time. Secondly, to rebut the update test objection, the authors stress that Flow Fragmentalism features fundamentally tensed facts and Priorian operators, which can account for robust passage. But one of the (arguably correct) criticisms of Iaquinto and Torrengo's towards standard tensed theories, such as those that attempt to explain passage in terms of Priorian operators, is that they cannot fully explain passage. Yet here they seem to appeal to standard resources to undermine the Update Test. Moreover, in standard tensed theories, tensed facts obtain in an absolute way, whereas in Flow Fragmentalism they obtain only relative to a fragment. This aspect risks making the whole picture static, since if a tensed fact obtains relative to a fragment *f*, it will always be the case that it obtains relative to fragment *f*.

4. Conclusion

In conclusion, *Fragmenting Reality* by Iaquinto and Torrengo is a highly insightful and rigorously argued work that provides a novel perspective on the metaphysics of time. The authors successfully offer a clear and precise account of time's passage while engaging with both A-theoretic and B-theoretic theories. Scholars working in the metaphysics of time, as well as those interested in causation and related topics, will find this book a valuable and thought-provoking contribution that advances our understanding. Its clarity, rigor, and innovative approach make the book an essential reading for anyone engaged in these areas of philosophy.

Acknowledgements

This research was funded in whole by the Austrian Science Fund (FWF) [grant DOI 10.55776/ESP196]. A translation in Italian of this review will appear in the journal APhEx (www.aphex.it).

References

Bernstein. Sara. 2017. "Time Travel and the Movable Present." In *Being, Freedom, and Method: Themes from the Philosophy of Peter van Inwagen,* edited by John A. Keller, 80–94. https://doi.org/10.1093/acprof:oso/9780198715702.003.0005

Effingham. Nikk. 2021. Vacillating Time: A Metaphysics for Time Travel and Geachianism. *Synthese*, 199(3): 7159–80. https://doi.org/10.1007/s11229-021-03108-5

Fine. Kit. 2005. “Tense and Reality.” In *Modality and Tense: Philosophical Papers*, edited by Kit Fine, 261–320. Oxford: Oxford University Press. https://doi.org/10.1093/0199278709.003.0009

Fine. K. (2006). The Reality of Tense. *Synthese*, 150(3): 399–414. https://doi.org/10.1007/s11229-005-5515-8

Hudson, Hud, and Ryan Wasserman. 2010. “Van Inwagen on Time Travel and Changing the Past.” in *Oxford Studies in Metaphysics*, vol. 5, edited by Dean Zimmermann, 41–49. Oxford: Oxford University Press.

Contents

RESEARCH ARTICLES

DISCUSSION NOTE

BOOK REVIEW

REPORTS

ADDENDUM