

Contents

Research Articles

Jonathan Bartlett: <i>Causal Capabilities of Teleology and Teleonomy in Life and Evolution</i>	222
Serhiy Kiš: <i>Does Deep Moral Disagreement Exist in Real Life?</i>	255
Alex Blum: <i>On Everything Is Necessarily What It Is</i>	278
Madelaine Angelova-Elchinova: <i>Perfect Thinkers, Perfect Speakers and Internalism about Thought Content</i>	281

Causal Capabilities of Teleology and Teleonomy in Life and Evolution

Jonathan Bartlett*


Received: 6 February 2023 / Revised: 3 July 2023 / Accepted: 28 July 2023

Abstract: Teleological causes have been generally disfavored in biological explanations because they have been thought to lack rigor or act as stand-ins for non-teleological processes which are simply not yet understood sufficiently. Teleological explanations in biology have been limited to only teleonomic causes, which are teleological causes that are due to codes or similarly reified mechanisms. However, advances in the conceptualization of teleological and teleonomic causation have allowed for more quantitative analyses of both. Additionally, although teleonomy has been historically excluded from potential causes of evolution, new research has shown that teleonomy actually plays a significant role in evolution. Combining these advances with advances in computability theory and information theory have allowed for a more rigorous and quantitative analysis of the capabilities and limitations of teleonomy in evolution.

Keywords: Teleology; teleonomy; evolution; information theory; active information; causation.

* The Blyth Institute

 <https://orcid.org/0000-0002-5961-0327>

 The Blyth Institute, The Blyth Institute, Broken Arrow, Tulsa, OK, USA

 jonathan.l.bartlett@gmail.com



1. Introduction: teleonomy and teleology

Teleology is the direction of causal outcomes of an entity (usually an organism) towards a purpose. Teleological concepts in biology (and especially in evolution) have had a troubling history for the simple reason that teleology tends to allow a lot of handwaving regarding causation (Hanke 2004; Turner 2017). Teleology has often been viewed by biologists (especially in the 19th and early 20th centuries) as being a temporary stand-in for more rigorous (i.e., reductive) ways of analyzing nature (Morgan 1905; Hanke 2004; LeMaster 2017; Crawford 2020). Biology in the twentieth century was therefore largely restricted to reductive forms of causation. This limitation allowed for tremendous growth in understanding certain *parts* of the organism, but it tended to leave out the organism itself (Woese 2004). Ultimately, while the twentieth century yielded large advancements in understanding the *mechanisms* of organisms, the results failed to include many defining traits of organisms, such as purpose and desire (Turner 2017).

Historically, biology has included teleology in one of two ways—externalist teleology (where teleology *outside* the organism plays a defining role) or internalist teleology (where the focus is on the purposes of the organism itself). Interestingly, while modern biology generally excludes external teleology in its causal toolkit, it can now be found as an undercurrent within physics in the question of cosmological fine-tuning (Barrow and Tipler 1998). While less favored in biology, discussions about biological fine-tuning have started to appear (Carr and Rees 2003; Bialek and Setayeshgar 2005; Thorvaldsen and Hössjer 2020), sometimes even connecting an externalist teleological framework of physics to the emergence of life as we know it (Morris 2004; Barrow et al. 2007). More recently, others have suggested external “teleological fields” operating at several levels that combine to produce teleological results (Babcock and McShea 2021).

Biologists, however, tend to favor an internalist teleology, where the focus is on the organism itself, focusing on aspects of causation such as intention and choice (Kull 2022). This is sometimes even connected to evolution (Fodor and Piatelli-Palmarini 2010; Kull 2022). In this understanding, biological language such as “selection for” are not merely artifacts of

language that are used accidentally, but represent some form of purpose within biological evolution.

Additionally, the present manuscript makes a distinction between “primary teleology,” where teleology is a first-class (irreducible) cause in the system, and “general teleology” which includes both primary teleology and also teleonomy, which is purposeful behavior which is due to a code or mechanism.¹ Usually, biologists will favor appeals to teleonomy over primary teleology, as teleonomic explanations do not require anything beyond physico-chemical explanation (Mayr 1992). Teleonomy was originally thought to allow for goal-directed processes but without requiring any connection to primary teleology—either internalist or externalist.

To illustrate these types of teleology, a non-exhaustive list of examples is shown in Table 1. The goal of this list is to not say which of these types of causes are real (the present author does not subscribe to all of them) but to give a clear picture on how various types of causes would be categorized under this taxonomy of teleology.

This paper has three primary goals. The first is to further develop Mayr’s teleonomy as a useful concept for biology to account for mechanistic aspects of purposeful biological causation without necessarily reducing the entire organism to a mechanism. The second is to show how recent advancements in information theory can be used to build an operationalized view of teleology that can be practically measured within organisms. The third goal is to combine these concepts to show new ways that the internal teleology of organisms can be helpfully embedded within a larger evolutionary framework.

2. The usefulness and distinctiveness of teleonomy

Recently, some have argued for changing the definition of teleonomy to cover all of internal teleology, without distinguishing primary teleology and teleology that has been reified into a mechanism (Corning 2019). The idea

¹ When the word “teleology” is used without a modifier in the present manuscript it will be considered as a shorthand for “general teleology.” Additionally, unless otherwise specified, “teleonomy” will refer to teleonomy in internal teleology.

is that Mayr's view of biology is too tied to a gene-centered, reductionist view of biology, where the organism is an inert entity that evolution happens to, rather than a participant in the process. However, there is nothing in Mayr's definition of "teleonomy" that ties one to this approach. In fact, having a *distinction* between teleonomy and primary internal teleology actually allows one to talk intelligibly about the differences between causes that are the result of true intentionality (primary internal teleology) and causes that are the result of intentionality reified in mechanism or code (teleonomy). If one conflates the two definitions, then this important distinction is lost. One can easily reject Mayr's conclusions *about* teleonomy while embracing the distinction and definition itself.

Others, such as Nagel (1977), have argued that there is not a clear distinction between Mayr's concepts of teleonomic processes (processes that exhibit ends due to a code) and teleomatic processes (processes which exhibit ends due to physics). An example of a teleomatic process would be a baseball being thrown. The end is the target that the ball is being thrown at, but the moving force here is physics. The reason that some view these as identical is that, in both cases (teleonomic and teleomatic processes), physics is essentially the only process under consideration, and therefore could be considered the "cause" of the outcomes in both cases. However, further analysis shows that teleomatic and teleonomic processes really do have distinct causal patterns.

Let us start by considering a process which is neither teleonomic nor teleomatic—processes for which no distinct end exists. In such a process, no end may exist either because the process is open (there are too many degrees of freedom so that it is indeterminate from initial conditions) or because the process is chaotic (while the outcome may be deterministic, it is not determinable through computable means). A teleonomic process is similar, except that a code or other control mechanism exists which controls the outcome based on logical principles, either reducing the inherent degrees of freedom of the physical process or making an unpredictable process predictable or understandable from logic, even if it is not tractable using equations. To illustrate this distinction, consider the difference between a traditional bomb dropped from an airplane and a smart bomb. A traditional bomb uses entirely teleomatic processes in-flight. The physics of the system is what

guides the outcome. A smart bomb, however, can use a control system to make adjustments in-flight according to the logic of what it is trying to do, such as follow a heat source. While the initial conditions *determine* the outcome, the outcome is likely not tractably predictable from mere mechanical considerations. However, someone understanding the logic of the control system would be able to largely predict the outcome to a high degree of accuracy.

To understand why this is the case, it is important to understand the nature of code and software systems. Software is usually written using Turing-complete programming languages. In such languages, in most cases it is impossible to mechanically predict any non-trivial property of the system in the general case (Rice 1953). Additionally, this unpredictability is not specifically the result of sophistication in the programming system, but is available even in extremely trivial cellular automata (Cook 2004). However, a program provides a control mechanism which makes the outcome predictable by the logic of what the program's outcome is intending. In the example of the smart bomb, if we know what the code on the bomb is tracking, we can predict where it will hit, even if it is not tractably determinable by mechanical modeling. So, with both teleomatic and teleonomic processes, initial conditions largely determine the outcome. However, with teleonomic processes, the prediction of that outcome requires understanding the logical structure of the control process, which would not be possible (or at least tractable) by direct analysis of the physical quantities.²

In any case, a teleological process is one whose causes are organized around purposes, and a teleonomic process is a teleological process that has been reified into a code or a mechanism. Keeping teleonomy as a distinctive category will allow us to analyze the limitations of teleonomy, which are not coextensive with the limitations of general teleology, as teleonomy is also limited by the fact that it must be reified into mechanism or code, and thus will inherit additional limitations from that reification.

² As a note, there are processes which are teleonomic but which are also tractable, such as if someone created a very simple code or mechanism to accomplish a goal. In such a case, someone would be able to understand the system in both teleonomic and teleomatic ways, as you could understand what was happening either from understanding the logic of the code or from modeling the states of the physical system.

3. Mathematically modeling non-mechanistic causation

The biggest hurdle in analyzing primary teleology has been a lack of mathematical frameworks for doing so. While mathematics may not be able to incorporate all that is included within the concept of primary teleology, there is no reason to think that there are no aspects of it which allow for mathematicization. This is especially true if we do not arbitrarily restrict our mathematical concepts to those that are computable.

While the definition of what counts as “mechanical,” “materialistic,” or “physical” has varied considerably over the centuries (Bartlett 2017c), many people today draw a rough equivalency between computability and material causation (Wolfram 2002; van Rooij 2008; Bartlett 2014; Copeland and Shagrir 2020).³ This equivalency goes by many names, including “the principal of computational equivalence,” “computationalism,” “the physical Church-Turing thesis,” and the “tractable cognition thesis.” Computationalism is naturally considered by many to be associated with mechanical causation because computation includes all results achievable by finitary processes. Thus, to go beyond what is achievable computationally would require non-finitary action in the world, which many view as being beyond what is meant by the idea of mechanism.

Even if one does not fully adhere to this equivalency, it at least provides a less ambiguous starting point for helping to understand and model primary teleological processes as distinct from mechanical processes, with primary teleological processes being those which are not directly computable. Computation, however, is only one aspect of mathematics, even if it is the most well-known. Mathematics has grown beyond solely being about computation and computable problems, especially since the 1930s with Kurt Gödel and Alan Turing. Gödel’s “incompleteness theorems” (Gödel 1931) demonstrated that one could prove mathematically the existence of

³ Here, “material causation” refers to what is also generally known as “naturalism,” not to the Aristotelian notion of material causes. However, using the term “naturalism” to refer to a specific understanding of metaphysics seems to be prejudicial, as it implies that any cause stated within that metaphysical paradigm is “natural,” and therefore any other cause would be “unnatural,” prejudicially implying that one would not expect to encounter such causes on an ordinary basis.

mathematical statements that were non-provable within their own axiomatic framework. While initially Gödel's result did not turn many heads, the application of it by Alan Turing's seminal paper (Turing 1937) started to show that Gödel's theorem had merit.

Turing, while inventing the concept of a computer program, was able to prove that we cannot write a computer program which will tell whether or not another arbitrary computer program will ever halt. For context, though the word often has negative connotations, "halting" is generally considered good in computer science—halting means that the computational task actually completes and has a result, as opposed to getting, say, an endlessly spinning cursor waiting for a computation to finish that never does. The fact that we cannot write a computer program which will tell us if another arbitrary program will halt is known in computer science as the "halting problem." This result has been further generalized to Rice's theorem, which says that any non-trivial property of a computer program cannot be proven using computation in the general case (Rice 1953). Given the Church-Turing thesis that Turing machines effectively exhaust what can be deduced through computation (Kripke 2013), this presents quite a limit on the abilities of computation (and therefore material causation) alone.

However, humans do not appear to be so limited. Humans can do a variety of tasks which have appeared to many mathematicians and computer scientists to be beyond the boundary of computation, such as determining whether or not computer programs will halt, determining new mathematical axioms, formalizing propositions, and deriving non-trivial properties of computer programs (Robertson 1999; Bringsjord 1997; Bringsjord and Zenzen 2003; Bringsjord and Arkoudas 2004; Bringsjord et al. 2006; Bartlett 2014).

Even though there are tasks which are beyond the reach of computation, there is no problem reasoning about such tasks mathematically, nor problems reasoning about entities which are capable of those tasks. Shortly after describing the limits of computation, Turing (1939) demonstrated the ability to mathematically reason about non-computable tasks using a concept now called "Turing oracles," which are essentially non-computable functions. The lack of being able to directly compute the *value* of such functions does not prevent the ability to *reason* about them (Copeland 1998; Bartlett 2014).

As noted in prior work (Bartlett 2017a), using such functions in models does bring some additional (but not insurmountable) problems to testability. Testability always plays an important role in science because it allows reality to push back against our ideas. Since values cannot be precomputed for such functions, other means of testing have to be applied. For instance, we might be able to predict the frequency of occurrence, or some parameter of the distribution of effects, or even a qualitative aspect of the distribution of effects.

While this doesn't test every aspect we might wish (after all, we wish we could know the value of the function ahead-of-time), complete empirical testability has never been absolute for any model in science. Empirical testing is by its nature more limited than the theories that it tests, as there are actually an infinite number of models which match *any* given set of data (Kukla 1996). In fact, the very existence of *p*-values tells us that there is *some* probability that the empirical testing was insufficient. Testing does not prove the validity of theories, it is merely a means to give voice to external reality about the content of theories.

The testing of and for randomness already shows that this sort of testing is practiced in science. Whether a finite set came from a random process cannot be determined by any finite set of data. However, for the purposes of testing, oftentimes randomness is determined by checking to see if the anticipated statistical parameters of the dataset match the expected distribution. For instance, in Luria-Delbrück experiments, the data are presumed to be following a Poisson distribution if the mean is equal to the variance (Luria and Delbrück 1943), despite the fact that there are an infinite number of ways that the mean can equal the variance without following a Poisson distribution.

The main criteria for an empirical test are that it is able to compare the consistency of *some* empirical parameter to one that is expected from the theory. This parameter does not have to be the specific value obtained, but can also be meta-information about the values, the preconditions of achieving them, the means of achieving them, their distribution, etc.

In short, primary teleological causes can be thought of as non-algorithmic functions, and can also be thought about and reasoned about as such.

4. Measuring teleological causation

One of the key limitations of current teleological thinking is the lack of measurements for the amount of general teleology in a process (Lee and McShea 2020). So how might teleology be measured?

In many recent expositions (Koons 1998; Hitchcock 1996; Hawthorne and Nolan 2006), teleology is identified with shifting probabilities towards some occurrence ϕ when (a) ϕ brings about result ψ and (b) ψ is good. This is illustrated conceptually in Figure 1. This shifting of probabilities based on outcomes is, essentially, a reduction or alteration of possibility space towards beneficial outcomes. Within the constraints of physics, an organism has many potential options. Teleology reduces those options to one particular value (or a smaller subset of values), or shifts probabilities in their favor, in service to a goal or holistic form (Asma 1996). If one needs a more objective way of determining a goal in biological systems, one can use the criteria established by Mossio and Bich (2017), where the goal is at least an aspect of the maintenance of the state of closure of the teleological system (though there are likely other ways to determine goals in an objective manner as well).

These reductions are typically “surprising” within the context of the physical system. That is, while statistical entropy encourages us to rely on systems achieving their most probable end-state, organisms tend to make choices and create configurations that are highly improbable statistically. This is sometimes referred to as the “cybernetic cut” that exists between outcomes whose causes are dominated by non-teleological causes and teleological causes (Abel 2008; Trevors and Abel 2004). This reduction in possibility space to an improbable but goal-conforming outcome is a key marker for teleological activity.⁴

⁴ Note that many of the attempts to reconcile or reduce all teleological causation into non-teleological processes such as natural selection have so far failed to accomplish their goals. As pointed out by Fodor and Piatelli-Palmarini (2010), this reconciliation, as it has been performed so far, only works if selection itself is teleological instead of strictly material. Block and Kitcher (2010) criticize this view, but merely by assertion. They state that “causation is extensional,” but their only support is that they can provide an *example* of a cause which is extensional. The fact that a

What makes teleology special is that the solution space tends to be extremely small within the possibility space. Therefore, we can measure the “power” of the teleological action by measuring this reduction in space. This is not always easy or exact because what has to be measured is the size of the destination space which matches the desired goals. We cannot presume that there is only one possible outcome that fits the goal, even though in biology (as well as in computation) the solution space is usually quite miniscule compared to the possibility space (Abel and Trevors 2005; Langdon 2006; Montañez 2017).

Because these solution spaces represent a tiny fraction of a possibility space, the probability of them occurring by chance are usually measured in bits for convenience, which is the negative log (base 2) of the probability (Marks II, Dembski, and Ewert 2016). This also allows for more convenient manipulation of values, as well as more intuitively-connected ways of manipulating information measurements (i.e., information is sub-additive, and therefore “adding information” is truly analogous to addition in this scheme). Thus, using the concept of bits from information theory provides a convenient methodology for representing reductions from possibility space to solution space in teleology.

Therefore, the amount of information that is added to a system to direct its outcome can be quantified by examining the change in probabilities that occurs. One way this can be measured is using the concept of active information (Dembski and Marks II 2009). Active information measures the difference in probability of finding a solution in a general space I_{Ω} (such as

particular cause is extensional does nothing to say whether or not there exists causes which are intensional, yet this is precisely what these critics do. Then they use this to simply rule that “selection for” must be an extensional cause, saying, “But if causation is extensional, then so is selection-for, since selection-for is a causal idea.” Mossio and Bich (2017) do a better job by describing the requirements for a physical process to be understood as being teleological. While I am in agreement with their assessment (such processes would fall under the category of teleonomy in the taxonomy of teleology presented here), their assessment does not include a reason to think that *all* teleological processes can be implemented entirely by physical processes, nor that such teleological processes are reasonably likely to *originate* in the prior total absence of a primary teleological process.

among physical possibilities) and finding a solution in a space I_S that is informed by an information source. The difference, I_+ , represents active information.

$$I_+ = I_\Omega - I_S \quad (1)$$

In terms of direct probabilities, this can be reformulated in an equivalent manner using p_Ω as the probability of success in the general space and p_S as the probability that is informed by an information source.

$$I_+ = -\log_2 \left(\frac{p_\Omega}{p_S} \right) \quad (2)$$

The concept of active information does not specify a source for the information, or even how the information is stored, it only measures its effects.⁵ Active information is typically used to measure the amount of information present in various algorithms, such as in Ewert, Dembski, and Marks II (2009) and Montañez et al. (2010), but it has also been used in measuring the degree of fine tuning in cosmology (Díaz-Pachón, Hössjer, and Marks II 2021) and judging the effectiveness of machine learning models (Bartlett and Holloway 2019). The fact that it is being designated as “information” does not mean that it is stored in an explicit digital format (it would be difficult to even conceptualize a “digital format” for cosmological fine-tuning, for instance), but merely that it can be reasoned about using information theory.⁶

⁵ For understanding the information content itself from an “inside the information” perspective, the reader is referred to the subject of teleosemantics.

⁶ While there are many who object to the overuse of information theory in biology, I believe such criticisms do not hold here. Using information theory for analyzing genetics itself is relatively uncontroversial, whether it is the maintenance of genetic information through time (Kuruoglu and Arndt 2017), the calculation of the entropy of the genetic code (Yockey 2000), or the measurements of the genetic “address space” provided by binding sequence lengths (Schneider et al. 1986). Any wider usage of information theory tends to be criticized as taking the information concept too far, and assuming that the entirety of causal factors for organisms reduce to computational ones (Griffiths 2001). These are not completely unfounded criticisms, as many improper or overly-encompassing analogies to computer systems have been

4.1. Comparison with the persistence measurement

Recently, Lee and McShea (2020) developed an empirical measurement of goal-directedness that they labeled as *persistence* which has similar goals as the previous concept of active information. This section will show that while the persistence metric heads in the right direction to some extent, active information appears to be a mathematically superior way of measuring goal directedness using equivalent inputs.

In their formulation of persistence,

$$P = \frac{\frac{G}{N} - R}{1 - R} \quad (3)$$

where P is the persistence measurement, $\frac{G}{N}$ is the ratio of good moves to total moves, and R is the expected ratio given the probability structure of the space. For any ratio $\frac{G}{N} \geq R$, since $G \leq N$ and $0 \leq R < 1$, the result will be in the range $0 \leq P \leq 1$.

made throughout the years. However, information theory, in its more abstract incarnations, applies much more widely than most assume, and does not even necessarily rely on computational metaphors. Demirel (2014) and Griffiths et al. (2015) both provide good examples of applying information theory without relying on computational metaphors.

The present discussion focuses on information theory in two ways. First, as a measurement tool, using active information. As mentioned already, active information does not rely on any computational metaphor to be usable for measurement, and has already shown its usefulness both inside and outside of computational systems. Second, as a source of providing limiting behavior or conservation rules for teleonomic systems. Here, the analogy to computation is more direct, but its application is more limited. That is, we are not applying information theory to the whole of the biological organism, but only to parts which can be determined to follow computable rules as delineated in the physical Church-Turing thesis. Information theory provides both the necessary requirements for inclusion in this analysis as well as the analysis tools themselves. For aspects of biological organisms which do not fit these requirements, the application of these limitations simply does not apply. This is what makes the taxonomy of teleology described in Section 1 helpful—it provides a way of at least naming the type of causality being proposed and therefore the tools relevant to its investigation.

To adequately compare persistence to active information, we need to establish some sort of equivalency between the terms in each formulation. For a first-order comparison of the measurements, we can use the equivalencies $I_\Omega = -\log_2(R)$ and $I_S = -\log_2\left(\frac{G}{N}\right)$ to bridge the two measurements.

While the measurements have some similarities, there are several advantages to active information over persistence. The first advantage is that, as an information measure, active information is sub-additive, while persistence is not. This means that active information measurements can be combined in meaningful ways. Being *sub*-additive means that adding information measures leads to an upper bound, not to a single value. Nonetheless, in the case of persistence, no means of combining values are provided at all. In Lee and McShea (2020), they suggest separating out different subspaces by probability structure. Separating such spaces in persistence measurements means that the results of the spaces are not combinable, while they would at least be sub-additive using active information.

The second advantage of active information is that the structure of the space generated by persistence does not make as much sense as that of active information. Both active information and persistence attempt to discount the teleological process when the probability structure dictates high success rates, and they both yield negative values when those processes point away from the goals. However, in both cases active information makes more mathematical sense than persistence, and it also provides additional structural benefits as well.

For instance, the negative side (where the “teleology” is actually pointing the wrong way) of persistence seems degenerate—the positive case yields values from 0 to 1, but the negative side can diverge to any negative value. If $\frac{G}{N} = 0.99$ and $R = 0.9$, the result is 0.9, but, if they are reversed, the result is -9 . Active information is symmetric on both the positive and negative sides. For the same probabilities, the active information is ≈ 0.095 bits, but, if $\frac{G}{N}$ and R were reversed, the active information is ≈ -0.095 bits. Lee and McShea (2020) actually considers all negative values to be zero, as all such values indicate that the organism is not oriented towards the goal. While this may be true, the lack of symmetry indicates that the measurement is not well-grounded mathematically.

As for discounting for the environmental assistance, while persistence does do some discounting, the discounting for persistence is neither sufficiently high nor does it track well. For instance, if the process performs 100% good moves, that isn't especially surprising if the environment dictates a probability space that is 99.99% favorable. Persistence, however, measures this as a 1 (the highest score). Active information, on the other hand, would only count that as ≈ 0.000144 bits. Where the environment entirely dictates the outcome, persistence is indeterminate while active information is zero, indicating that there is nothing being enhanced to the environment's distribution (persistence is simply undefined in this case). As noted in Griffiths et al. (2015), causal specificity is an important aspect of causation, and here active information is much better able to measure this specificity than persistence.⁷

Additionally, when considering degenerate cases where where success is literally impossible but success occurs anyway (i.e., R is 0 but $\frac{G}{N}$ is positive), active information correctly yields infinite values, while persistence simply yields $\frac{G}{N}$. It is difficult to imagine how we should not be impressed by the teleological accomplishment of the impossible, but here persistence is giving it an equal or lower score than for the easy accomplishment of 100% accuracy when the environment yields 99.99% assistance. Likewise, this continues to be problematic in non-degenerate cases where R is merely miniscule or infinitesimal where the limit is likewise $\frac{G}{N}$. Essentially, while persistence

⁷ In reference to Griffiths et al. (2015), I should make a note about the relationship of their measurement of causal specificity using mutual information and our measure of teleology using active information, as the two measurements are closely related. Mutual information essentially averages active information across all available possibilities, while active information focuses on the specific possibilities of interest (here, teleological goals). In other words, mutual information as used by Griffiths et al. (2015) measures total causal influence between cause and effect, while active information focuses on causes related to specific classes of effects (i.e., meeting a goal). Mutual information is always non-negative, so it would measure a cause pointing away from the goal as having positive mutual information, even though the information is in the wrong direction. Active information (which is a form of *pointwise* mutual information) allows for indicating both causal specificity *and* directionality.

can provide a limited ability to factor *out* the help the environment provides, it seems to fail completely to factor *in* the difficulty the environment provides.

4.2. How might primary teleology interact with physics?

One potential problem with any approach to teleology which includes primary teleology and distinguishes physical and teleological probability spaces (whether active information, persistence, or some other measurement) is that models and methodologies have not been established for the interactions of the physical and the teleological, leaving open the question of how I_S can come to be different from I_Ω or how $\frac{G}{N}$ can come to be different from R .⁸ Babcock and McShea (2021) have suggested that teleology operates as a field much like other fields that have been discovered. They note that fields “are multiply realizable and diverse in their composition,” and, thus, there is no fundamental issue at play. Whether one takes an internalist or externalist view to primary teleology, the same question arises—how does one think about the interaction of the teleological and the non-teleological? While it is an open question (and not directly addressed by the present methods), it is certainly not *problematic* for the present methodologies.

An example means of addressing such a question would be to take the physical possible outcomes and the teleological possible outcomes and, with the intersection of their possibilities, combine their probabilities in some fashion and then remap the probabilities so that they add up to 1. For instance, let Ω be an array of possible future states of the universe, and let α be an array of the probabilities of each of those states where the index of α matches the index of Ω that it is standing in for. Now let β be a similar array but for the probabilities based solely on primary teleology. We can

⁸ This question is not as directly relevant when dealing with teleonomic causes, because the information or control system manages the change in probability space, as noted in Section 2. However, even then, a more ultimate question remains about how those control systems came to be, and, as will be discussed in Section 7, the degree of difficulty gets larger, not smaller, as it gets pushed back in time, thus indicating that, at some point, primary teleology will likely be required.

combine these distributions into a new array, γ , by performing the operation

$$\gamma_i = \frac{\alpha_i \beta_i}{\sum_j \alpha_j \beta_j}$$

with potentially some additional failsafes to prevent an undefined distribution.⁹ The point here is not that this particular suggestion is the correct model (or even an approximation of a correct model), but merely to demonstrate that there is not a conceptual problem with having models that distinguish primary teleological and physical probability spaces.

5. What requires teleological causation?

To begin a quantitative investigation into teleological processes (whether primarily teleological or teleonomical), we first need to be able to identify them qualitatively. What we need to know are what sort of outcomes require teleological processes to occur.

There are many processes which might occur either through teleological means or through non-teleological means. A rock could tumble down a mountain either because the wind blew it down or because someone pushed it. There is usually nothing in the nature of the tumbling rock which would give a clue as to which type of cause occurred. However, because teleological solutions *can* narrow the solution space by many orders of magnitude, we can in fact detect certain types of solutions which require (within a miniscule margin of error) teleological processes.

It is insufficient, though, to simply measure differences in probabilities of outcomes given certain starting points. In order for something to qualify as being teleological, it has to be identifiable as having a goal. For the purposes of biology, one can postulate that a goal is something that an organism can do or cause to happen which benefits the organism itself in some way. This is categorized as an “intrinsic purpose” by Koons (1998). In biology, a more objective way of identifying a goal is given by Mossio

⁹ For instance, if there is no overlap between physical and teleological possibility space, all terms become $\frac{0}{0}$ under this formalism.

and Bich (2017), where the goal should be something involved with the maintenance of the closure of the organism. However, benefit itself is insufficient to establish something as teleological, as there may be processes which are beneficial but also simply expected from the environment. Instead, one needs to demonstrate in some way that an operation is sufficiently more likely to occur when beneficial than when not. That is, if the end-goal modifies the probability space to a sufficient degree to lean towards the goal, then we have justification to call the organism's actions teleological.

The degree of probability-space modification required for justifying attributing a process to teleological causes is determined by convention just like p -value or α -level thresholds are determined by convention for justifying material causes. This is a new area of inquiry, so official conventions have not been established. Since active information is a specified complexity model (Bartlett 2020a) and α -levels in such models can be converted to bits ($-\log_2(\alpha)$) for comparison (Montañez 2018), we can use α -level conventions for significance tests. This would yield 4.33 bits for an α -level of 0.05, or 6.65 bits for an α -level of 0.01. These represent the low end of what might qualify for recognizing an event as teleological.

On the other end of the spectrum is what is often referred to as the “universal probability” bound. This is a probability limit which, when used in specified complexity models, render the specified outcome not just unlikely, but that it would go beyond exhausting the probability resources of the whole history of the universe to achieve the result (Dembski 2006; Abel 2009). Depending on the source, this has been variously calculated as being between 360 and 500 bits.

Unfortunately, 4.33–500 bits is a huge range for which conventions have not been established. However, keeping in mind that the upper end of this range is only there to provide near-certainty across all time and space, I would offer 10 bits as generally being good evidence for teleological behavior, as it is an order of magnitude beyond the conventional requirements for inferring material causation.

6. Limitations of teleonomy compared to primary teleology

Teleonomy is essentially teleology that is due to a code or program (Mayr 1961). In the present theoretical framework, teleonomy represents prior, existing information that can be leveraged by an organism in order to accomplish a goal. In other words, this code provides information to the organism about likely ways that possibility space can be reduced to solution space. The information need not be total—partial reductions in possibility space are helpful as long as they reduce it to within an amount that allows for an organism to reasonably find a solution.

Therefore, teleonomy can function similarly to primary teleology within this framework. However, teleonomy can only function this way within a bounded (though possibly large) scope. Once outside this scope, teleonomy no longer supplies information to the process, and in fact can potentially detract from the process if the boundaries of the usefulness of the code do not match the organism's present circumstances.

Teleonomy does not have to be a code per se, but any sort of mechanism (as defined in Section 3) suffices for teleonomic considerations. However, teleonomy occurring through a code is both easier to analyze (as the possibility space is easier to examine qualitatively and quantitatively) and it is present (and required) for evolvable systems, as certain types of reproduction require code-based reproduction (Mignea 2014).¹⁰ Therefore, although teleonomy does not require codes per se, our analyses will tend to be code-focused for simplicity, straightforwardness, and applicability to reproducing evolvable systems. However, information theory can be applied to any biophysical system through entropy analysis, as shown in Demirel (2014). Processes involving codes merely make the process much more tractable in common cases.

¹⁰ A simple explanation of why reproduction requires a code is to imagine a copier of physical structures compared to a code-based copier. For many physical structures, the actual process to achieve the end-result is not inferable from the structure itself. Therefore, the “thing” that is copied has to be some form of information, not the thing itself.

Interestingly, because we are analyzing code-based teleonomy, many of its properties can be investigated using computability theory.¹¹ We can use computability theory to determine the limits of what the teleonomic codes *themselves* are capable of. This does not necessarily limit the capabilities of the *organism*, as it may have other sources of teleology (primary internal teleology or general external teleology). But we can find limits of what internal teleonomy alone may be able to accomplish within an organism.

One interesting insight is that many tasks that require primary teleology for a general solution can be supplied a specific or semi-specific solution using code. For an abstract example, take the halting problem discussed previously. While the halting problem is not solvable for programs *generally*, code can be added to solve for specific instances or classes of the halting problem. That is, in code, we could write detection code to determine if, say, a loop counter did not go in the proper direction in order to reach a termination condition. Likewise, we could write detection code that detected when the same state was achieved more than once (which necessarily indicates an infinite loop in deterministic systems).

However, what is generally not possible is teleonomy (code) that produces new code that performs outside the classes of solution that it already considers.¹² This is often known as Levin's law, and it is colloquially stated as "torturing an uninformed witness cannot give information about the crime" (Levin 1984). This result is stable in deterministic, non-

¹¹ Against the objection that biological codes are unlike computer codes, here we are only considering computability theory in general, which finds general truths applicable to all coded systems, regardless of the specifics of the way that the codes are mapped onto function. Additionally, for those considering material causes to be coextensive with computation as described in Section 3, this would apply to all material causes whether or not we conceptualized the underlying system as a "code."

¹² It is possible for someone to separately *identify* additional classes of solutions. The problem, however, is that an organism (or program) does not encounter new classes of challenges based on their appropriateness to in-built solutions, but rather as the environment presents them. Therefore, while there will likely be additional classes of problems that the code can solve, the proviso "generally not possible" refers to the fact that organisms are encountering challenges that are from outside, and therefore independent of what the organism might have programming to solve.

deterministic, and mixed contexts, and is sometimes termed as “conservation of information” (Holloway 2020).

Note that, unlike other conservation laws, conservation of information is not absolute. Information content (measured in bits) *can* grow; it just grows extremely slowly. Since informational bits are the negative log (base 2) of probability, information content can only grow with the *logarithm* of the number of probabilistic attempts at a solution (see Figure 2). In this case, while geological time can provide some amount of information, the universal probability metric (which takes into account the estimated age and size of the universe) limits the amount of information gained to 500 bits as previously noted. Indeed, as noted in Hössjer, Bechly, and Gauger (2021), without imposing external sources of information, waiting times increase exponentially with the size of the needed mutation.

So, while teleonomy can be used to contain a relatively static amount of information towards organismal teleology, it does have limits based on the initial information content. These limits cannot be stated *a priori* because, for any finite set of mechanistic (i.e., computable) challenges, there could be code that assisted the organism in solving or mitigating the challenge. Some such mechanisms in biology have been reviewed in Payne and Wagner (2019). However, because the amount of code or number of mechanisms itself is finite, there will be potential challenges whose solutions are unattainable strictly through teleonomy.

7. Teleonomy and evolution

Evolutionary theory has often excluded teleological causes historically. However, the recent advent of evolutionary teleonomy (Corning 2014; Bartlett 2017b) has started to reincorporate teleonomic causes into the process of evolution. Additionally, recent advances have shown that the evidences that were previously used to exclude teleonomic understandings of evolutionary processes were misinterpreted (Bartlett 2023).

Within evolution, the role of teleonomy and teleology generally to produce evolutionary novelty effectively occupy the same basic roles as for producing goal-oriented solutions during an organism’s life, with the important difference that, here, the target is the genetic code itself. Effectively,

evolution is acting (at least mathematically) as a search for a solution within code to solve a problem. Therefore, since it is operating mathematically as a search, that means that the mathematics of search apply to the production of evolution, whether or not evolution is ontologically a search (which is a matter of debate). When dealing with teleonomy, this leads to a conservation of information problem that mirrors the one in the previous section. That is, we cannot add significant amounts of information by simply processing what is already there.

However, what can be done is for an organism to have information about what sorts of external environments require what sort of internal changes. Essentially, organisms maintain not only the explicit genome, but what Caporale calls an “implicit genome,” which consists of not only the present state of the genome but also of the states which it is programmed to achieve (Caporale 2006). Teleonomy can maintain a partial mapping of external environments and potentially successful DNA configurations, or even just a mapping of likely biologically valid operations on existing DNA. This has been observed repeatedly by many investigators at least as far back as the 1990s (Hall 1999), many of them reviewed in Caporale (2006), Zhang et al. (2013), Bartlett (2017c), and Bartlett (2020b).

Stochasticity does play a role here. However, because the number of possible configurations grows exponentially with the size of a search space, teleonomy is required in order to bring the number of attempted states down to a reasonable level that selection can sort through on timescales required by the population (see Figure 3). Such teleonomy in evolution can be measured using the techniques described in Section 4. Bartlett (2020b) provides several equations that can be used in this measurement along with examples of it being applied.

One example of its application is in measuring the amount of teleology *E. coli* demonstrates when adapting to starvation in the presence of citrus, as explored in Hofwegen et al. (2016). They found that, when under selection, *E. coli* produced *Cit+* mutations faster than when not under selection. When under selection, getting a *Cit+* mutation is needed for maintaining self-closure, so it qualifies as a teleological goal. To quantify the teleology using active information, we would compare the probability that the mutation occurred when not under selection to the probability that the mutation

occurred when under selection, showing that *E. coli* contributes ≈ 12.4 bits of information towards the search for *Cit+* mutations when under selection.

This was measured by using the probability of a mutation to yield a *Cit+* phenotype when occurring in absence of selection, which is $\frac{1}{10^{10}}$, compared to the probability of a mutation to yield a *Cit+* phenotype when occurring in the presence of selection (i.e., when needed), which is $\frac{1}{1.8 \times 10^6}$. Therefore, the active information metric is calculated from (2) as $I_+ = -\log_2 \left(\frac{1.8 \times 10^6}{10^{10}} \right) \approx 12.4$ bits. This is greater than the threshold proposed in Section 5, and indicates that this process is indeed teleological. We can go further with some additional simplifying assumptions (which may or may not turn out to be true). If we assume that the teleology is teleonomic, and that the teleonomy is codified within DNA, then we can expect that there is a mechanism encoded within the DNA which is causing these mutations to be more likely when under selection. However, active information is not equivalent to data length, so the amount of DNA required to accomplish this cannot be inferred from active information alone. Thus, as a practical benefit, active information can be used to help know when the search for a mutational mechanism is justified, as such experiments can have significant associated costs.

Many, such as Caporale (2006), suppose that, with time, evolution will be able to add to its ability to find solutions using material mechanisms alone (e.g., the mechanisms provided by the modern synthesis, or some other material mechanism). Since selection favors the ability to find solutions, it is supposed that organisms will evolve an increased ability to find solutions over time. However, the mathematics of search says that this is not likely. In the mathematics of searching, evolving a more evolvable system would be equivalent to a search for a search. However, a search for a search actually requires *more* initial teleonomy than the search itself, not less. This result has been formalized in search as the “displacement theorem” (Dembski and Marks II 2010). Essentially, this means that teleonomy only provides for bounded or parameterized amounts of novelty, which is in accord with what is known from information theory (Holloway and Marks II 2018). This is backed up by empirical studies of the general capabilities of evolutionary algorithms (Woodward and Bai 2009; Yampolskiy 2018).

One common objection to the above idea is that the information could be in the environment rather than the organism. This is technically true, but doesn't actually solve the problem. The reduction of possibility space to solution space is a massive reduction, and performing the reduction requires specific information which matches the structure of the *organism*. In order for the organism to take advantage of this, one of two things must be true. Either the organism already knows the range of variation in the environment, and is already coded to search and match it (in which case we are back to the information actually residing in the organism), or the environment has specific information on how to modify the organism to match the environment. In the latter case, this merely reduces to an externalized teleonomy.

This is much like a factory which contains information about how the objects it produces should turn out. It is true that the information is external to the objects being produced, but it does not get around the need for teleonomy in the process as a whole, it just locates it externally.

Note that these limits are not the limits of evolution per se, only the limits of teleonomically-directed evolution (and, by extension, evolution without general teleology). There is nothing here which excludes primary teleology, whether internal or external, for which limitations are not known. Even though the limits of teleonomy are not equivalent with the limits of evolution, making such a distinction is important because recognizing the fundamental limiting principles of mechanical types of causes help us to locate, discover, and analyzes those types of causes. It does not imply a limitation on the total range of causes available, just a way to analyze those that are encompassed by the prerequisites of teleonomy.

8. Conclusion

As we have seen, we can generate a more mathematical conception of teleology by looking at it from the perspective of probability spaces and information theory. The mathematical form of teleology is the ability of an organism to reduce the possibilities from the ones provided by physics to ones that are more likely to be in accord with the organism's own goals. This reduction, or at least aspects of it, is measurable.

Teleonomy is essentially encoded teleology. As such, information theory and computability theory provide abstract tools that enable further investigation of teleonomy’s capabilities and limitations. Teleonomy, while historically not applied to evolution, has been shown in recent years to have a much more important role in the evolutionary process. However, the limitations imposed by information and computability theory give a limitation to how much teleonomy can contribute to evolution without relying on additional teleological causes.

Figures and Tables

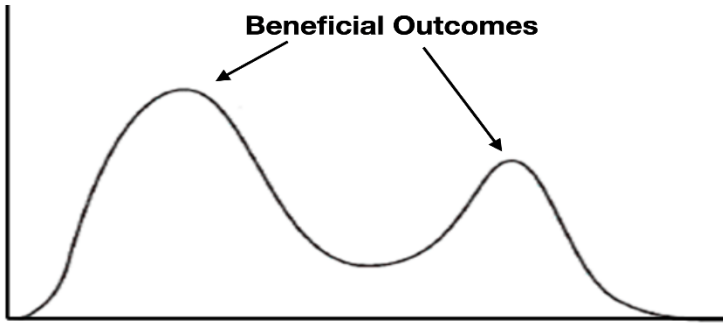
Table 1: Examples of Types of Teleology

	Internal Teleology	External Teleology
Primary Teleology	hylomorphic form, soul, free will	other organisms with primary teleology, cosmic teleology (such as teleology inherent in nature or the universe, deity, etc.)
Teleonomy	biophysical control systems, developmental processes, DNA, histone code, sugar code	cultural rules and expectations, physical laws of form (see Denton and Marshall (2001))

Figure 1: Teleological Causes Shift Distributions: Conceptual Illustration

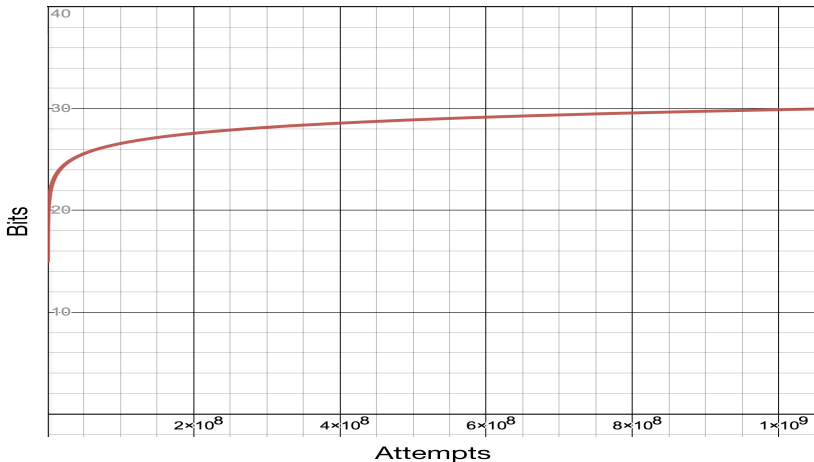


The graphic above represents a distribution of possible outcomes as might be expected from physics.



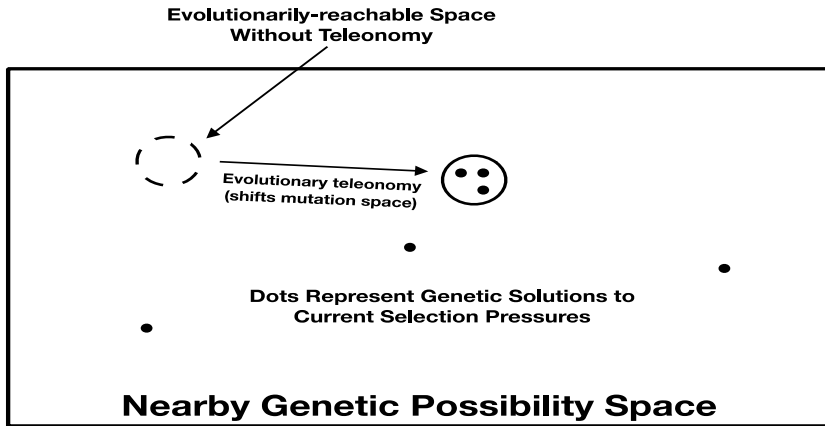
The graphic above represents a shifting of probabilities of outcomes due to teleology, where beneficial outcomes are now favored. The possibility space can remain the same even as the probabilities of outcomes change. The teleology of these shifts can be quantified using active information.

Figure 2: Information Generation Under Conservation of Information



The maximum number of bits of information content that can be generated from teleonomic sources according to conservation of information is the log (base 2) of the number of attempts. As pictured, even with 10^9 attempts, not even 30 bits of information content can be generated. Note that information content here is not necessarily equivalent to code or data size.

Figure 3: A Depiction of Evolutionary Teleonomy



Although the possible genetic space that a population can search is large, it is miniscule compared to even the nearby possibility space (orders of magnitude more miniscule than depicted here). The dashed circle represents the size and location of genetic search space of a population by unassisted mutations (in the modern synthesis, selection does not change local mutation space). Evolutionary teleonomy represents the shifting of the mutation space for the population so that individuals are more likely to hit targets. As shown, not all possible targets need to be included in this new possibility space, only more targets than were included in the prior probability space.

References

- Abel, David L. 2008. "The 'Cybernetic Cut': Progressing from Description to Prescription in Systems Theory." *The Open Cybernetics & Systemics Journal* 2 (1): 252–62. <https://doi.org/10.2174/1874110X00802010252>.
- Abel, David L. 2009. "The Universal Probability Metric (UPM) & Principle." *Theoretical Biology and Medical Modelling* 6 (27). <https://doi.org/10.1186/1742-4682-6-27>.
- Abel, David L, and Jack T Trevors. 2005. "Three Subsets of Sequence Complexity and Their Relevance to Biopolymeric Information." *Theoretical Biology and Medical Modelling* 2 (1): 1–16. <https://doi.org/10.1186/1742-4682-2-29>.

- Asma, Stephen T. 1996. "A Neglected Teleology." In *Following Form and Function: A Philosophical Archaeology of Life Science*, 128–46. Northwestern University Press.
- Babcock, Gunnar, and Daniel W McShea. 2021. "An Externalist Teleology." *Synthese* 199: 8755–80. <https://doi.org/10.1007/s11229-021-03181-w>.
- Barrow, John D, Simon C Morris, Stephen J Freeland, and Charles L Harper Jr, eds. 2007. *Fitness of the Cosmos for Life*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511536557>.
- Barrow, John D, and Frank J Tipler. 1998. *The Anthropic Cosmological Principle*. Oxford University Press.
- Bartlett, Jonathan L. 2014. "Using Turing Oracles in Cognitive Models of Problem-Solving." In *Engineering and the Ultimate: An Interdisciplinary Investigation of Order and Design in Nature and Craft*, edited by Jonathan L Bartlett, D Halsmer, and M R Hall, 99–122. Blyth Institute Press. <https://doi.org/10.33014/isbn.0975283863.5>.
- Bartlett, Jonathan L. 2017a. "Describable but Not Predictable: Mathematical Modeling and Non-Naturalistic Causation." In *Naturalism and Its Alternatives in Scientific Methodologies*, edited by Jonathan L Bartlett and Eric M Holloway, 113–27. Blyth Institute Press. <https://doi.org/10.33014/isbn.1944918078.7>.
- Bartlett, Jonathan L. 2017b. "Evolutionary Teleonomy as a Unifying Principle for the Extended Evolutionary Synthesis." *BIO-Complexity* 2017 (2): 1–7. <https://doi.org/10.5048/BIO-C.2017.2>.
- Bartlett, Jonathan L. 2017c. "Philosophical Shortcomings of Methodological Naturalism and the Path Forward." In *Naturalism and Its Alternatives in Scientific Methodologies*, edited by Jonathan L Bartlett and Eric M Holloway, 13–38. Blyth Institute Press. <https://doi.org/10.33014/isbn.1944918078.2>.
- Bartlett, Jonathan L. 2020a. "Active Information Is a Specified Complexity Model." *Communications of the Blyth Institute* 2 (2): 40–41. <https://doi.org/10.33014/issn.2640-5652.2.2.bartlett.1>.
- Bartlett, Jonathan L. 2020b. "Measuring Active Information in Biological Systems." *BIO-Complexity* 2020 (2): 1–11. <https://doi.org/10.5048/BIO-C.2020.2>.
- Bartlett, Jonathan L. 2023. "Random with Respect to Fitness or External Selection? An Important but Often Overlooked Distinction." *Acta Biotheoretica* 71 (2). <https://doi.org/10.1007/s10441-023-09464-8>.
- Bartlett, Jonathan L, and Eric M Holloway. 2019. "Generalized Information: A Straightforward Method for Judging Machine Learning Models." *Communications of the Blyth Institute* 1 (2): 13–21. <https://doi.org/10.33014/issn.2640-5652.1.2.bartlett.1>.

- Bialek, William, and Sima Setayeshgar. 2005. "Physical Limits to Biological Signaling." *Proceedings of the National Academy of Sciences* 102 (29): 10040–45. <https://doi.org/10.1073/pnas.0504321102>.
- Block, Ned, and Philip Kitcher. 2010. "Misunderstanding Darwin." *Boston Review*.
- Bringsjord, Selmer. 1997. "An Argument for the Uncomputability of Infinitary Mathematical Expertise." In *Expertise in Context*, edited by P Feltovich, K Ford, and P Hayes. AAAI Press.
- Bringsjord, Selmer, and Konstantine Arkoudas. 2004. "The Modal Argument for Hypercomputing Minds." *Theoretical Computer Science* 317: 167–90. <https://doi.org/10.1016/j.tcs.2003.12.010>.
- Bringsjord, Selmer, Owen Kellett, Andrew Shilliday, Joshua Taylor, Bram van Heuveln, Yingrui Yan, Jeffrey Baumes, and Kyle Ross. 2006. "A New Gödelian Argument for Hypercomputing Minds Based on the Busy Beaver Problem." *Applied Mathematics and Computation* 176 (2): 516–30. <https://doi.org/10.1016/j.amc.2005.09.071>.
- Bringsjord, Selmer, and M Zenzen. 2003. *Superminds: People Harness Hypercomputation, and More*. Kluwer Academic Publishers.
- Caporale, Lynn H, ed. 2006. *The Implicit Genome*. Oxford University Press.
- Carr, B J, and M J Rees. 2003. "Fine-Tuning in Living Systems." *International Journal of Astrobiology* 2 (2): 79–86. <https://doi.org/10.1017/S1473550403001472>.
- Cook, Matthew. 2004. "Universality in Elementary Cellular Automata." *Complex Systems* 15 (1): 1–40.
- Copeland, B J. 1998. "Turing's O-Machines, Searle, Penrose, and the Brain." *Analysis* 58 (2). <https://doi.org/10.1111/1467-8284.00113>.
- Copeland, B J, and Oron Shagrir. 2020. "Physical Computability Theses." In *Quantum, Probability, Logic*, 217–31. Springer. https://doi.org/10.1007/978-3-030-34316-3_9.
- Corning, Peter A. 2014. "Evolution 'on Purpose': How Behaviour Has Shaped the Evolutionary Process." *Biological Journal of the Linnean Society* 112 (2): 242–60. <https://doi.org/10.1111/bij.12061>.
- Corning, Peter A. 2019. "Teleonomy and the Proximate–Ultimate Distinction Revisited." *Biological Journal of the Linnean Society* 127 (4): 912–16. <https://doi.org/10.1093/biolinnean/blz087>.
- Crawford, Annie L. 2020. "Metaphor and Meaning in the Teleological Language of Biology." *Communications of the Blyth Institute* 2 (2): 5–24. <https://doi.org/10.33014/issn.2640-5652.2.2.crawford.1>.
- Dembski, William A. 2006. *The Design Inference: Eliminating Chance Through Small Probabilities*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511570643>.

- Dembski, William A, and Robert J Marks II. 2009. "Conservation of Information in Search: Measuring the Cost of Success." *IEEE Transactions on Systems, Man and Cybernetics A, Systems & Humans* 5 (5): 1051–61. <https://doi.org/10.1109/TSMCA.2009.2025027>.
- Dembski, William A, and Robert J Marks II. 2010. "The Search for a Search: Measuring the Information Cost of Higher Level Search." *Journal of Advanced Computational Intelligence and Intelligent Informatics* 14 (5): 475–86. <https://doi.org/10.20965/jaciii.2010.p0475>.
- Demirel, Yaşar. 2014. "Information in Biological Systems and the Fluctuation Theorem." *Entropy* 16: 1931–48. <https://doi.org/10.3390/e16041931>.
- Denton, Michael, and Craig Marshall. 2001. "Laws of Form Revisited." *Nature* 410: 417. <https://doi.org/10.1038/35068645>.
- Díaz-Pachón, Daniel A, Ola Hössjer, and Robert J Marks II. 2021. "Is Cosmological Tuning Fine or Coarse?" *Journal of Cosmology and Astroparticle Physics* 2021 (20). <https://doi.org/10.1088/1475-7516/2021/07/020>.
- Ewert, Winston, William A Dembski, and Robert J Marks II. 2009. "Evolutionary Synthesis of Nand Logic: Dissecting a Digital Organism." In *Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics*, 3047–53. IEEE. <https://doi.org/10.1109/ICSMC.2009.5345941>.
- Fodor, Jerry, and Massimo Piatelli-Palmarini. 2010. *What Darwin Got Wrong*. Farrar, Straus; Giroux.
- Gödel, Kurt. 1931. "On Formally Undecidable Propositions of Principia Mathematica and Related Systems I." *Monatshefte Für Mathematik Und Physik* 38: 173–98. <https://doi.org/10.1007/BF01700692>.
- Griffiths, Paul E. 2001. "Genetic Information: A Metaphor in Search of a Theory." *Philosophy of Science* 68 (3): 394–412. <https://doi.org/10.1086/392891>.
- Griffiths, Paul E, Arnaud Pocheville, Brett Calcott, Karola Stotz, Hyunju Kim, and Rob Knight. 2015. "Measuring Causal Specificity." *Philosophy of Science* 82 (4): 529–55. <https://doi.org/10.1086/682914>.
- Hall, Barry G. 1999. "Transposable Elements as Activators of Cryptic Genes in *E. Coli*." *Genetica* 107: 181–87. <https://doi.org/10.1023/A:1003936706129>.
- Hanke, David. 2004. "Teleology: The Explanation That Bedevils Biology." In *Explanations: Styles of Explanation in Science*, edited by John Cornwell, 143–55. Oxford University Press.
- Hawthorne, John, and Daniel Nolan. 2006. "What Would Teleological Causation Be?" In *Metaphysical Essays*, edited by John Hawthorne. Clarendon Press. <https://doi.org/10.1093/acprof:oso/9780199291236.003.0015>.
- Hitchcock, Christopher. 1996. "A Probabilistic Theory of Second Order Causation." *Erkenntnis* 44 (3): 369–77. <https://doi.org/10.1007/BF00167664>.

- Hofwegen, Dustin J Van, Carolyn J Hovde, Scott A Minnich, and T J Silhavy. 2016. "Rapid Evolution of Citrate Utilization by *Escherichia Coli* by Direct Selection Requires *citT* and *dctA*." *Journal of Bacteriology* 198 (7): 1022–34. <https://doi.org/10.1128/JB.00831-15>.
- Holloway, Eric M. 2020. "Independence Conservation and Evolutionary Algorithms." *Communications of the Blyth Institute* 2 (1): 32–35. <https://doi.org/10.33014/issn.2640-5652.2.1.holloway.2>.
- Holloway, Eric M, and Robert J Marks II. 2018. "Observation of Unbounded Novelty in Evolutionary Algorithms Is Unknowable." In *ICAISC 2018: Artificial Intelligence and Soft Computing*, edited by L Rutkowski, R Scherer, M Korytkowski, W Pedrycz, R Tadeusiewicz, and J M Zurada, 395–404. Springer. https://doi.org/10.1007/978-3-319-91253-0_37.
- Hössjer, Ola, Gunther Bechly, and Ann K Gauger. 2021. "On the Waiting Time Until Coordinated Mutations Get Fixed in Regulatory Sequences." *Journal of Theoretical Biology* 524 (110657). <https://doi.org/10.1016/j.jtbi.2021.110657>.
- Koons, Robert C. 1998. "Teleology as Higher-Order Causation." *Minds and Machines* 8: 559–85.
- Kripke, Saul A. 2013. "The Church-Turing 'Thesis' as a Special Corollary of Gödel's Incompleteness Theorem." In *Computability: Turing, Gödel, Church, and Beyond*. MIT Press.
- Kukla, Andre. 1996. "Does Every Theory Have Empirically-Equivalent Rivals?" *Erkenntnis* 44: 137–66. <https://doi.org/10.1007/BF00166499>.
- Kull, Kalevi. 2022. "Choices by Organisms: On the Role of Freedom in Behaviour and Evolution." *Biological Journal of the Linnean Society*. <https://doi.org/10.1093/biolinnean/blac077>.
- Kuruoglu, Ercan E, and Peter F Arndt. 2017. "The Information Capacity of the Genetic Code: Is the Natural Code Optimal?" *Journal of Theoretical Biology* 419: 227–37. <https://doi.org/10.1016/j.jtbi.2017.01.046>.
- Langdon, William B. 2006. "The Halting Probability in Von Neumann Architectures." *Lecture Notes in Computer Science* 3905: 225–37. https://doi.org/10.1007/11729976_20.
- Lee, Jong Gwan, and Daniel W McShea. 2020. "Operationalizing Goal Directedness: An Empirical Route to Advancing a Philosophical Discussion." *Philosophy, Theory, and Practice in Biology* 12 (5): 1–31. <https://doi.org/10.3998/ptpbio.16039257.0012.005>.
- LeMaster, James C. 2017. "The Relationship of Bacon, Teleology, and Analogy to the Doctrine of Methodological Naturalism." In *Naturalism and Its Alternatives in Scientific Methodologies*, 67–89. Blyth Institute Press.

- Levin, Leonid A. 1984. "Randomness Conservation Inequalities; Information and Independence in Mathematical Theories." *Information and Control* 61: 15–37. [https://doi.org/10.1016/S0019-9958\(84\)80060-1](https://doi.org/10.1016/S0019-9958(84)80060-1).
- Luria, S E, and M Delbrück. 1943. "Mutations of Bacteria from Virus Sensitivity to Virus Resistance." *Genetics* 28 (6): 491–511. <https://doi.org/10.1093/genetics/28.6.491>.
- Marks II, Robert J, William A Dembski, and Winston Ewert. 2016. "Conservation of Information in Computer Search." In *Introduction to Evolutionary Informatics*, 105–86. World Scientific.
- Mayr, Ernst. 1961. "Cause and Effect in Biology." *Science* 134: 1501–6. <https://doi.org/10.1126/science.134.3489.1501>.
- Mayr, Ernst. 1992. "The Idea of Teleology." *Journal of the History of Ideas* 53: 117–35. <https://doi.org/10.2307/2709913>.
- Mignea, Arminius. 2014. "Developing Insights into the Design of the Simplest Self-Replicator and Its Complexity: Part 2—Evaluating the Complexity of a Concrete Implementation of an Artificial SSR." In *Engineering and the Ultimate: An Interdisciplinary Investigation of Order and Design in Nature and Craft*, edited by Jonathan L Bartlett, Dominic Halsmer, and Mark R Hall, 187–212. Blyth Institute Press. <https://doi.org/10.33014/isbn.0975283863.10>.
- Montañez, George D. 2017. "The Famine of Forte: Few Search Problems Greatly Favor Your Algorithm." In *IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 477–82. IEEE. <https://doi.org/10.1109/SMC.2017.8122651>.
- Montañez, George D. 2018. "A Unified Model of Complex Specified Information." *BIO-Complexity* 2018 (4): 1–26. <https://doi.org/10.5048/BIO-C.2018.4>.
- Montañez, George D, Winston Ewert, William A Dembski, and Robert J Marks II. 2010. "A Vivisection of the Ev Computer Organism: Identifying Sources of Active Information." *BIO-Complexity* 2010 (3): 1–7. <https://doi.org/10.5048/BIO-C.2010.3>.
- Morgan, Conway Lloyd. 1905. *The Interpretation of Nature*. Macmillan.
- Morris, Simon C. 2004. *Life's Solution: Inevitable Humans in a Lonely Universe*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511535499>.
- Mossio, Matteo, and Leonardo Bich. 2017. "What Makes Biological Organisms Teleological?" *Synthese* 194: 1089–1114. <https://doi.org/10.1007/s11229-014-0594-z>.
- Nagel, Ernst. 1977. "Goal-Directed Processes in Biology." *The Journal of Philosophy* 74 (5): 261–79. <https://doi.org/10.2307/2025745>.
- Payne, Joshua L, and Andreas Wagner. 2019. "The Causes of Evolvability and Their Evolution." *Nature Reviews Genetics* 20: 24–38. <https://doi.org/10.1038/s41576-018-0069-z>.

- Rice, H G. 1953. "Classes of Recursively Enumerable Sets and Their Decision Problems." *Transactions of the American Mathematical Society* 74 (2): 358–66. <https://doi.org/10.2307/1990888>.
- Robertson, D S. 1999. "Algorithmic Information Theory, Free Will, and the Turing Test." *Complexity* 4 (3): 25–34. [https://doi.org/10.1002/\(SICI\)1099-0526\(199901/02\)4:3<25::AID-CPLX5>3.0.CO;2-E](https://doi.org/10.1002/(SICI)1099-0526(199901/02)4:3<25::AID-CPLX5>3.0.CO;2-E).
- Schneider, Thomas D, Gary D Stormo, Larry Gold, and Andrzej Ehrenfeucht. 1986. "Information Content of Binding Sites on Nucleotide Sequences." *Journal of Molecular Biology* 188 (3): 415–31. [https://doi.org/10.1016/0022-2836\(86\)90165-8](https://doi.org/10.1016/0022-2836(86)90165-8).
- Thorvaldsen, Steinar, and Ola Hössjer. 2020. "Using Statistical Methods to Model the Fine-Tuning of Molecular Machines and Systems." *Journal of Theoretical Biology* 501. <https://doi.org/10.1016/j.jtbi.2020.110352>.
- Trevors, Jack T, and David L Abel. 2004. "Chance and Necessity Do Not Explain the Origin of Life." *Cell Biology International* 28 (11): 729–39. <https://doi.org/10.1016/j.cellbi.2004.06.006>.
- Turing, Alan M. 1937. "On Computable Numbers, with an Application to the Entscheidungsproblem." *Proceedings of the London Mathematical Society* s2-42 (1): 230–65. <https://doi.org/10.1112/plms/s2-42.1.230>.
- Turing, Alan M. 1939. "Systems of Logic Based on Ordinals." *Proceedings of the London Mathematical Society* s2-45: 161–228. <https://doi.org/10.1112/plms/s2-45.1.161>.
- Turner, J S. 2017. *Purpose and Desire*. HarperOne.
- van Rooij, Iris. 2008. "The Tractable Cognition Thesis." *Cognitive Science: A Multidisciplinary Journal* 32 (6). <https://doi.org/10.1080/03640210801897856>.
- Woese, Carl R. 2004. "A New Biology for a New Century." *Microbiology and Molecular Biology Reviews* 68 (2): 173–86. <https://doi.org/10.1128/MMBR.68.2.173-186.2004>.
- Wolfram, Stephen. 2002. *A New Kind of Science*. Wolfram Media.
- Woodward, John R, and Ruibin Bai. 2009. "Why Evolution Is Not a Good Paradigm for Program Induction: A Critique of Genetic Programming." In *Proceedings of the First ACM/SIGEVO Summit on Genetic and Evolutionary Computation*, 593–600. Association for Computing Machinery. <https://doi.org/10.1145/1543834.1543915>.
- Yampolskiy, Roman V. 2018. "Why Do We Not Evolve Software? Analysis of Evolutionary Algorithms." *Evolutionary Bioinformatics* 14: 1–11. <https://doi.org/10.1177/1176934318815906>.
- Yockey, Hubert. 2000. "Origin of Life on Earth and Shannon's Theory of Communication." *Computers and Chemistry* 24: 105–23. [https://doi.org/10.1016/S0097-8485\(99\)00050-9](https://doi.org/10.1016/S0097-8485(99)00050-9).

Zhang, Zongghe, Jing Wang, Maksim A Shlykov, and Milton H Saier Jr. 2013. “Transposon Mutagenesis in Disease, Drug Discovery, and Bacterial Evolution.” In *Stress-Induced Mutagenesis*, edited by D Mittelman, 59–77. Springer. https://doi.org/10.1007/978-1-4614-6280-4_4.

Does Deep Moral Disagreement Exist in Real Life?

Serhiy Kiš*

Received: 13 February 2023 / Revised: 3 July 2023 / Accepted: 26 August 2023

Abstract: The existence of deep moral disagreement is used in support of views ranging from moral relativism to the impossibility of moral expertise. This is done despite the fact that it is not at all clear whether deep moral disagreements actually occur, as the usually given examples are never of real life situations, but of some generalized debates on controversial issues. The paper will try to remedy this, as any strength of arguments appealing to deep moral disagreement is partly depended on the fact the disagreement exists. This will be done by showing that some real life conflicts that are intractable, i.e. notoriously difficult to resolve, share some important features with deep moral disagreement. The article also deals with the objection that the mere conceptual possibility renders illustrations of actually happening deep moral disagreements unnecessary. The problem with such objection is that it depends on theoretical assumptions (i.e. denial of moral realism) that are not uncontroversial. Instead, the article claims we need not only suppose deep moral disagreements exist because they actually occur when some intractable conflicts occur. Thus, in so far as to the deep moral disagreement's existence, the arguments appealing to it are safe. But as intractable conflicts can be resolved, by seeing deep moral disagreements as constitutive part of them, we might have to consider whether deep moral

* University of Pardubice

 <https://orcid.org/0000-0001-5031-4364>

 Department of Philosophy and Religious Studies, University of Pardubice, Stavařov 97 532 10 Pardubice, Czech Republic

 sergejkish@seznam.cz



disagreements are resolvable too. A brief suggestion of how that might look like is given in the end of the paper.

Keywords: Argument from disagreement; Deep moral disagreement; Intractable conflict; Israeli-Palestine conflict; South African apartheid

Introduction

“Deep” or “radical” moral disagreement is essentially a situation, in which parties hold incompatible moral values or principles and are thus unable to agree on a given moral issue. Significantly, what makes such disagreement deep is the impossibility to determine who in the disagreement is mistaken. The usual suspects such as fallacious reasoning, cognitive bias, or some other deficiency in the involved parties are out of the picture, as neither of the equally able “reasoners” (often called *epistemic peers*) suffer from it.

In some sub-disciplines of moral philosophy, such as moral epistemology or meta-ethics, this deep moral disagreement is often used in different kinds of arguments, ranging from support of moral relativism to denial of the notion of moral expertise. Further, any such argument presumably draws its strength partly from the fact that such deep moral disagreement exists. Oddly enough, arguments appealing to deep moral disagreements never quote particular instances of deep moral disagreements that *actually happened*. All that is given are intuitive, but nonetheless very general examples, such as debate on moral permissibility of abortion, homosexuality or meat-eating.

The usual explanation behind this is that deep moral disagreement need not actually exist, as the possibility of its existence validates arguments appealing to it well enough (Tolhurst 1987). This reasoning, however, is seriously challenged by objectivist replies, according to which deep moral disagreements do not exist and are also conceptually impossible (Parfit 2011).

The omission of *historical examples*, as I define them in a moment, thus merits skepticism about the existence of deep moral disagreements, and in turn arguments build on it. In this paper, I wish to put such suspicions to rest.

In what follows, I will show that deep moral disagreements do, in fact, exist, although not exactly in the way philosophical literature defines them. I will claim that the best real life instance of deep moral disagreement can be found in some “intractable conflicts” studied by the discipline of *peace and conflict studies*. Intractable conflict is long-lasting, mentally, and sometimes even physically destructive disagreement, which resists every attempt at a successful resolution (Deutsch, Coleman, and Marcus 2006). There are many forms of intractable conflicts and not all have deep moral disagreements in them. Many, however, do and this makes them fine historical examples of deep moral disagreements. My goal, then, is to show the salient features that both concepts share. This will be accomplished in the following manner.

First, I introduce the concept of deep moral disagreement. Second, I provide some illustrations of the arguments appealing to it and mention the fact that these arguments have serious sociopolitical consequences, making it much more important to prove the validity of their starting premise. Whenever possible, I will be bringing attention to the fact that the usual examples of deep moral disagreements are never of actual events, but instead of general moral debates that seem like they are deep. Third, I deal with a possible objection according to which it is unnecessary to prove the actual existence of deep moral disagreement, as the mere possibility validates the conclusions drawn from it. I rebut with objectivist arguments that deny even the mere possibility of deep moral disagreements. Finally, I introduce the concept of intractable conflict and, by exposing its salient features, argue that some of them are the best real life instances of deep moral disagreement. If I am right, philosophers need not worry about whether deep moral disagreements exist, or are possible. However, they will have to lessen their expectations regarding the irresolvability of deep moral disagreements.

1. Deep moral disagreement

In the most recent comprehensive review of the topic of disagreement in general, Ronald Rowland (2021) goes through all popular arguments from disagreement in the fields of (moral) epistemology, meta-ethics, normative

ethics, and political philosophy. Using his review, I want to define deep moral disagreement and start exposing the systematic lack of what I in a moment define as historic examples.

Disagreement, technically speaking, is a situation in which one person believes p and the other $\sim p$. An illustration of this, on Rowland's view, is two people disagreeing on what type of taxation policy helps the least well-off in the most effective way. For him, such disagreement is a matter of "non-moral empirical facts," and it can be settled as such (Rowland 2020, 2). In contrast, *moral* disagreements "would survive even if parties to these disagreements agreed on all the relevant non-moral facts and information" (Rowland 2021, 5). With regards to taxation, the parties of moral disagreement do not disagree on what policy should be applied, but on whether it is *just* to apply any taxation policy at all. Moral disagreement boils down to disagreement about moral values and principles, such as individualist vs. collectivist forms of morality. A more formal definition would thus be: moral disagreement is a situation in which one person wants to act according to the moral principle or value m and the other according to some moral principle or value that is incompatible and incommensurable with m (cf. Kekes 1996, chap. 4). But what makes moral disagreement *deep*?

Further on, in the context of epistemology, Rowland considers deep disagreements – for clarity, call them deep *non-moral* disagreements. For non-moral disagreement to be deep, two conditions must be met: (i) parties of disagreement have different ways of assessing evidence and what even counts as evidence, that is, they have different epistemic principles; and (ii) there exists no further (meta)epistemic principle that would settle the disagreement of first-order epistemic principles (Rowland 2021, 116). Rowland, interestingly enough, gives a real-life illustration of this: the disagreement between "old earthers and young earthers." These parties do not even agree on what counts as evidence (the Bible vs. data from radioactive dating), and there is seemingly little they can do about it.

Next, Rowland presents the use of deep non-moral disagreement in epistemological theorizing. Namely, he considers what are the implications of deep non-moral disagreement for confidence of our beliefs. On the *conciliat*ionist view, (non-deep) disagreement serves as higher order evidence of a mistake, forcing the parties involved to lower the confidence of their

respective beliefs.¹ This is not so when disagreement is deep, in which case the *steadfast* view holds: “if we find ourselves in a deep disagreement with another about whether p , this does not give us reason to lower our confidence or suspend belief about whether p ,” because this disagreement is explained by the parties’ adherence to different epistemic principles and not by a reasoning mistake they made (Rowland 2021, 116). It is also worth noting that Rowland provides a number of illustrations here. I will return to this point later.

Finally, Rowland arrives at deep moral disagreement. It is worth quoting him at length.

There seem to be some important moral epistemological implications of this view. At least some, perhaps many, disagreements about the moral status of abortion and homosexuality are deep. For some people believe that abortion and homosexuality are wrong on biblical or religious grounds. Those who disagree and do not form moral beliefs on the basis of biblical or religious interpretation find themselves in a deep disagreement about the morality of homosexuality and abortion. So neither party to these deep moral disagreements have the justification of their moral beliefs defeated or undermined by these deep moral disagreements. (Rowland 2021, 117)

Let me give two comments to this. First, notice how Rowland translates the conditions of deep non-moral disagreement to the sphere of moral disagreements. He portrays a picture, in which some people believe abortion to be wrong on “religious grounds,” whereas others disagree, because they “do not form moral beliefs on the basis of biblical or religious interpretation.” Thus, they have different moral principles or values.

Furthermore, Rowland adds that because parties have different grounds for their respective moral beliefs, the disagreement they are involved in does not defeat or undermine their respective moral beliefs. For that to be the

¹ “If we should believe that there is a substantial division of opinion among our (approximate) epistemic peers regarding whether p , then, other things equal, we should suspend belief about whether p or significantly lower our confidence about whether p ” (Rowland 2021, 89).

case, to repeat, the parties would have to share the grounds for their moral beliefs and there is only one way for them to start sharing the grounds – they must appeal to some moral meta-principle that would determine which of their starting moral principles or values should be abandoned. If that happened, the disagreement would be explained by the fact that one of the parties made reasoning mistake. The problem is, for Rowland, that they are unable to appeal to some moral meta-principle, “since the parties to these disagreements disagree about which principles generate more reliable results” (Rowland 2020, 118). In other words, they do not and *cannot* share the grounds for their respective moral beliefs. Thus, the disagreement stands despite no one being wrong and there is nothing that can be done about it.

In my reading, then, for moral disagreement to be deep, two conditions must be met: (i) parties of disagreement hold incompatible and incommensurable moral principles or values and (ii) there exists no further way of settling the disagreement between moral principles or values.

My second comment concerns the examples of deep moral disagreements that Rowland gives: morality of abortion and homosexuality. After this, he gives one more, stating that “[s]ome disagreements between act-utilitarians and their opponents may also be deep disagreements” (Rowland 2021, 116). Are these good examples of deep moral disagreement?

First off, what makes an example of deep moral disagreement good? At first approximation, an example of any phenomenon should arguably allow us to *grasp* or *get the hand of* what is salient about the phenomenon. We can use closer characterization given by Timothy Williamson in his paper on role of examples in “armchair philosophizing”:

Examples are almost never described in complete detail; a mass of background must be taken for granted; it cannot all be explicitly stipulated. Many of the missing details are irrelevant to whatever philosophical matter are in play. (Williamson 2005, 6)

It is this “philosophical matter,” in my estimation, that is supposed to be conveyed by useful examples. But the question now is, what is a “philosophical matter” that is supposed to be conveyed by examples of deep moral disagreements? Or put differently, what do we want from examples of deep

moral disagreement? I believe it is three things: conditions (i), (ii) and the example to be describing *actually existing event* (or *existence condition* for short). When an example of deep moral disagreement meets all three of these “philosophical matters,” I shall call it a *historical example of deep moral disagreement*, or *historical example* for short. When an example meets only (i) and (ii), as is most often the case, I shall call it *ahistorical example of deep moral disagreement*, or *ahistorical example* for short.

Are Rowland’s examples of deep moral disagreement – morality of abortion and homosexuality – historical? Undoubtedly, there were and still are debates on such issues. But do these debates meet (i) and (ii) too? This, I claim, we cannot know until we analyze their real life instances. Granted, Rowland suggests that from all the disagreements on morality of abortion or homosexuality, only “some” and “perhaps many” are *deep*. But why does he think so? The answer lies, I suppose, in the fact that it is possible to *imagine* they meet (i), (ii) and existence condition at the same time. This answer, however, will not do because there are serious objectivist challenges to it that would first need to be dealt with. I shall spend more time on this in section 3. Before that, let me explain what existence of deep moral disagreements is supposed to imply (and why it matters) by showcasing two different arguments appealing it: one from meta-ethics and the second from moral epistemology.

2. Arguments from deep moral disagreement

Let me start the exposition of arguments from the deep moral disagreement in meta-ethics. Namely, with arguably the most famous argument against moral objectivism: John Mackie’s “argument from relativity.”

Mackie appeals to the existence of deep moral disagreement² in arguing against moral objectivism, a thesis that – to put it in one way – moral

² Granted, he uses different terminology. But his conclusion that one and the same thing can arouse “radically and *irresolvably*” (Mackie, 1991, 38, emphasis mine) different moral judgments in different people – i.e. people may “irresolvably” disagree about something – leads me to believe he refers to what I defined as deep moral disagreements. See also citations of him below.

values “are not part of the fabric of the world” (Mackie 1991, 15, see also 29-30). If the values were objective, the argument goes, we would see agreement in people’s “ways of life” more often. However, we do not observe much of this agreement. Mackie asks why.

One possible answer is to say that there are “very general basic principles which are recognized at least implicitly to some extent in all society,” and which then “married with differing concrete circumstances [...], will beget different specific moral rules” (Mackie 1991, 37). Essentially, the reply goes, people follow the same moral principles, and the only difference is in the way they are deployed in virtue of their circumstances. This reply, however, is not enough for Mackie, as he does not believe this is what actually happens in life: universal moral principles “are very far from constituting the whole of what is actually affirmed as basic in ordinary moral thought.” Instead, what (Mackie claims) actually happens in life is that

people judge that some things are good or right, and others are bad or wrong, not because [...] they exemplify some general principle for which widespread implicit acceptance could be claimed, but because something about those things arouses certain responses immediately in them. (Mackie 1991, 37-8)

The second possible answer as to why there is not much agreement according to Mackie is the claim that many disagreements can also be seen in science. Disagreements there, however, do not lead us to refuse objectivity. Mackie’s rejoinder is simple: when disagreements occur in the sciences, they can be easily explained by showing that some mistakes in the research process were made. However, Mackie continues, “it is hardly plausible to interpret [deep] moral disagreement in the same way,” where the “causal connection” is the exact opposite: people approve of monogamy because they participate in it, not vice versa (Mackie 1991, 36). The conclusions come before the hypotheses are formulated, so to say.

Thus, Mackie’s belief that irresolvable differences among societies and individuals are best explained by refusing the notion that moral values are objective and instead accepting the relativity of values. But notice how the whole of Mackie’s argumentation just assumes deep moral disagreements exist, or as he puts it: “[t]he argument from relativity has as its premiss the

well-known variation in moral codes” (Mackie 1991, 36, emphasis mine). However, what, if pressed, would Mackie cite as a source or evidence for this premise? Besides passing remark on moral code’s variation being a “fact of anthropology,” nothing else is said.

His example, if we may call it so, does arguably meet *existence condition*, but we do not know if “variation in moral codes” is a result of (i) and (ii) also. The mentioned “fact of anthropology” would help us to determine it. Unfortunately – and unsurprisingly –, Mackie does not quote any anthropological or ethnographic studies. Let me now move to another example.

In her contribution to *Oxford Studies in Metaethics* series, Sarah McGrath considers “whether and to what extent moral disagreement undermines moral knowledge” (McGrath 2008, 87). Despite using a different terminology, she has in mind what I am calling deep moral disagreement. She terms a belief that is a subject of deep moral disagreement a (capital) “CONTROVERSIAL”³ belief, defining it in a following way: “belief is CONTROVERSIAL if and only if it is denied by another person of whom it is true that: you have no more reason to think that he or she is in error than you are” (McGrath 2008, 91). This is, however, different from (lower case) “controversial” belief, which she defines as “hotly contested” questions, such as morality of death penalty, abortion⁴, meat-eating or charity-giving

³ She, not entirely helpfully, uses in her text “CONTROVERSIAL” in capitals as technical term and “controversial” in lower case as everyday adjective.

⁴ Connected to this is one rather anecdotal proof of my point, which merits attention. It is located in Nicholas Sturgeon’s 1994 paper where he problematizes the connection between moral disagreement and moral relativism. In analyzing “unsettleable issues,” i.e. deep moral disagreements, he searches for examples: “Consider an example Foot and Wong both give of an unsettleable issue, that of the permissibility of abortion.” He however hesitates to use it, claiming “this would not be my example, since I regard the permissibility of at least early abortions as quite settleable.” He nevertheless accepts it in the end: “but since I do admit unsettleable questions, let me use their example for the sake of discussion” (Sturgeon 1994, 94). It is as if it did not really matter what example of moral disagreement we use, as long as we *assume* it is deep. That is, as long as we assume (i) and (ii). The problem is, as I argue in section 3, we are not justified – at least not uncontroversially – in assuming the existence condition in equal manner.

(McGrath 2008, 92-3). It follows that “controversial beliefs” may be, but need not be, “CONTROVERSIAL.” That is, “hotly contested” questions of morality of abortion or meat-eating may be, but need not be, topics of deep moral disagreements. McGrath analyzes whether we can claim knowledge of those topics that are “CONTROVERSIAL.”

She proposes a claim: “If one’s belief that p is CONTROVERSIAL, then one does not know that p ” (McGrath, 2008, 91, emphasis mine). Most importantly, she provides an example. To illustrate CONTROVERSIAL belief, she first asks the reader to imagine that they disagree with a friend, but that the friend has made a mistake. Then she compares that to a different situation: “But *suppose* instead that you have no such reason to think that it is Alice who has made the mistake: as far you know [sic], it is just as likely that you are mistaken as she is,” in such case, McGrath concludes, we are not justified in claiming knowledge of the disputed proposition (McGrath 2008, 92, emphasis mine). Unsurprisingly and most importantly for the present study, she asks us only to *suppose* we are in deep moral disagreement. No such actual situation is described. Here, the example is *ahistorical* as it meets (i) and (ii) but not existence condition.

The existence condition would be met by showing that our actual disputes on the permissibility of abortion or meat-eating are CONTROVERSIAL. However, this McGrath does not take for granted: “It is of course much less clear that [controversial moral beliefs] are also CONTROVERSIAL” (McGrath 2008, 93). She proceeds by examining possible ways in which controversial beliefs could be also CONTROVERSIAL, concluding it to be *possible*, but only if we conceive the disagreeing parties as having “relatively wide background of shared moral beliefs,” which would suggest they are both equally likely to be right in a dispute on, say, the permissibility of abortion (McGrath 2008, 106).

In sum, McGrath in her consideration of deep moral disagreement and the possibility of moral knowledge does not establish that deep moral disagreements occur, but only that they could, conceptually, occur. Therefore, her examples of deep moral disagreement are also *ahistorical*.

Before scrutinizing the strategy according to which it suffices for the arguments appealing to deep moral disagreement to establish its conceptual possibility, let me first explain why the validity of such arguments is

important beyond mere intellectual reasons. The matter of fact is, that the conclusions drawn from arguments appealing to deep moral disagreement are serious and consequential for well-being of people. Returning to Mackie, consider his opinion on importance of whether moral values are objective: “It clearly matters for general philosophy. [...] [H]ow this issue is settled will affect the possibility of certain kinds of moral argument” (Mackie 1991, 25). Consider here the classical of whether and who can intervene in situations of human rights abuses. If moral relativism holds, it will be very difficult to defend intervention by appealing to cross-cultural values.

Alternatively, consider the case of McGrath, where deep moral disagreements would lead us to abandon the notion that in morality some people’s opinion is above others’ in virtue of their expertise (McGrath 2008, cf. 105-6). This is troubling, as such experts are whom we trust with mitigating societal polarization and what today is called “culture wars.”

In summary, I presented two arguments appealing to the existence of deep moral disagreement and showed why the validity of these arguments matter. Furthermore, in the process, I have been bringing attention to the fact that all given examples of deep moral disagreement are ahistorical: they are about moral principles or values (i), they are impossible to settle (ii), but none of them are instances of real life disputes (existence condition). This omission merits skepticism towards the validity of arguments appealing to deep moral disagreement. However, there is defense against this omission. Namely, the mere *possibility* of deep moral disagreement validates the conclusions drawn from it. I shall now turn to this defense.

3. Deep moral disagreement as conceptual possibility

The obvious reaction to my claim is as follows. There is actually no need to empirically establish the existence of deep moral disagreement, as its mere conceptual possibility suffices for arguments appealing to it. This strategy is deployed, for example, by William Tolhurst (1987). Let me illustrate his reasoning.

Tolhurst, in a way, combines Mackie and McGrath, as he argues for moral relativism not by denying the existence of objective values, but by arguing they are “epistemically inaccessible,” and holding them is thus

never justified (Tolhurst 1987, 611). He does so in a following manner. First, he poses two epistemic principles which state, in short, that people with equal or similar “epistemically relevant features” (i.e. epistemic peers) cannot disagree about “justified objective proposition” (Tolhurst 1987, 611-12).

Next, he adds to a premise that there are situations in which epistemic peers disagree about justified objective propositions. In other words, Tolhurst assumes deep moral disagreements exist. This premise, however, is inconsistent with two stipulated epistemic principles – either the parties of deep moral disagreement are not epistemic peers, or objective proposition they disagree about is not justified. Tolhurst concludes with the former: “no objective moral beliefs are justified” (Tolhurst 1987, 613). Thus, under my terminology, what Tolhurst did was to *assume* or *stipulate* the existence condition. How does he justify this move?

On Tolhurst’s view, his “argument does not require any empirical premises concerning the nature and extent of *actual* [deep moral] disagreements, only the assumption that certain sort of moral disagreement is ubiquitously possible” (Tolhurst 1987, 610, emphasis mine). In other words, there is no need to give historical examples in his argument, because its strength lies only in it being “deductively valid” and as long as its premises are “plausible,” the validity is secured (Tolhurst 1987, 610).

Is, then, my requirement of historical examples of deep moral disagreement in arguments appealing to it justified? If we can establish something is a conceptual possibility, the arguments we draw from it are, logically speaking, fine enough. After all, this is fairly common practice (not only) in philosophy.⁵ It thus seems that existence condition in historical examples is unnecessary, as its conceptual possibility does the job as well. The problem with this, however, is that conceptual possibility of existence condition is “framework-dependent” and as such can be seriously challenged. Let me explain.

Say, just like Mackie, we want to give the best explanation of prevailing differences in people’s moral principles and values. We wish, to put it differently, to explain why – contrary to the opinions of political philosophers

⁵ This practice is sometimes called “counterfactual reasoning” and it is for example very popular among so called *compatibilist* in free will debate. See Austin (1979) for a classical illustration.

such as Francis Fukuyama (2006) – there is now seemingly less, not more, *overlap* in people’s views on what is morally right. Using a framework of relativism, our claim will be akin to that of Mackie’s or Tolhurst’s. It need not be, however. Imagine we are not, in fact, using a theoretical framework of relativism, but instead some other. Would then the lack of overlap in people’s views on morality be still best explained by them constituting deep moral disagreement? Let me draw on Derek Parfit (2011), perhaps the staunchest defender of moral realism, to answer that question.

Parfit disagrees (pun intended) with the notion that our inability to agree on moral matters is best explained by the fact that they are deep moral disagreements because of his view on the nature of moral claims. On Parfit’s account, moral claims, such as “It is permissible to have an abortion” or “It is not right to eat meat,” are *propositions* not different from claims expressing physical laws, such as “The speed of sound is approx. 343 m/s” or “Applied force being constant, the smaller the area, the bigger the pressure.” This means moral claims can be either true or false in virtue of facts independent on us (Parfit 2011, 391). Were that not so, the notion of being mistaken and, in turn, being able to improve morally, would be completely incoherent (Parfit 2011, 395). They are however not incoherent. Let me illustrate.

Say you think that abortion is never morally permissible because you believe that what results from the combination of sperm and egg holds the same status as a full-fledged person. Next, you encounter Judith Jarvis Thomson’s (1971) argument, according to which fetus being a person does not outweigh mother’s right to her body. Assume, moreover, you are persuaded by her argument. What happens next? You start to regard your prior belief regarding abortion to be mistaken, as now you believe that at least early-stage abortion is morally permissible.

Thus being “wrong” or “mistaken” is not only very important in our everyday moral lives, but it is also not incoherent. From this it follows, on Parfit’s view, that moral claims must be true or false propositions.

But if moral claims are propositions, then in moral disagreement there must always be someone who is mistaken: “When our value judgments express beliefs, which might be either true or false, we can claim that one of two conflicting [moral] judgments must be mistaken” (Parfit 2011, 391).

Recall the definition of disagreement above – one party believes p , the other $\sim p$. The question now is, what moral claims being propositions imply for the conceptual possibility of deep moral disagreement, or of (i) and (ii)? If a person for some reason mistakenly adheres to some moral principles or values, can conditions that we must disagree on morality (i) and there is nothing we can do about it (ii) still hold? Obviously, condition (i) is untouched by moral claims being propositions. But not so with (ii).

On Parfit's account, an "asymmetry" between disagreeing parties is always possible to establish by referencing the reasons for why one of the disagreeing parties is more likely to have got it right. Or as he puts it:

Since I believe that these other people are mistaken, there is one asymmetry between us. But I cannot rationally have much confidence in my beliefs unless there seems to be some other asymmetry, which would explain why it is these other people, and not me, who have made mistakes. There are often, I believe, such other asymmetries. My main example here will be the person from whom, in several disagreements, I have learned most. Williams was the most brilliant British moral philosopher whom I have known. If there were no other asymmetries between us, I could not rationally believe that it was I, rather than Williams, who was more likely to be right. (Parfit 2011, 430)

In conclusion, then, if one was to explain people's inability to agree on many moral issues not with relativist background, but a realist one, one's conclusion would be completely different. Instead of invoking the notion of deep moral disagreement, which in realist terms is conceptual impossibility as (ii) can never hold, we would concede that there is not much agreement on many issues, but only because many people have faulty judgments. When asked, skeptically, what is our justification for saying that most people are wrong, we could cite the works of Jonathan Haidt (2013) or Eleanor-Gordon Smith (2019), who argue – convincingly, in my estimation – that it is not reasons and facts, but emotions what mainly affects people's opinions on moral issues. In emotional affect, unsurprisingly, it is very difficult to not make a mistake in rational judgment. We could thus explain prevailing

moral disagreements without for a moment assuming deep moral disagreement to be a conceptual possibility.

This is why I consider strategies akin to that of Tolhurst's suspicious. To show his strategy is not controversial, he would first need to defeat Parfit's account of the nature of moral claims (and accounts similar to it) and establish that deep moral disagreement is, in fact, conceptually possible. Until then, it is not safe to just assume (i), (ii) and existence condition hold together. In the final section, I wish to argue we need not only assume it, as historical examples of deep moral disagreements can be given in a form of "intractable conflicts."

4. Some intractable conflicts are historical examples of deep moral disagreements

Merriam-Webster defines (tractable) conflict, in a couple of ways: it is "competitive or opposing action of incompatibles; antagonistic state or action (as of divergent ideas, interests, or persons); mental struggle resulting from incompatible or opposing needs, drives, wishes, or external or internal demands." Lastly, to be in conflict could also simply mean "to fail to be in agreement or accord."

One of the most authoritative sources on the subject is *The Handbook of Conflict Resolution* (Deutsch, Coleman, and Marcus 2006). On its outset we are given a vivid description of real life conflict between husband and wife:

The destructiveness of their way of dealing with their conflicts was reflected in their tendency to escalate a dispute about almost any specific issue (e.g., a household chore, the child's bedtime) into a power struggle in which each spouse felt that his or her self-esteem or core identity was at stake. The destructive process resulted in (as well as from) justified mutual suspicion; correctly perceived mutual hostility; a win-lose orientation to their conflicts; a tendency to act so as to lead the other to respond in a way that would confirm one's worst suspicion; inability to understand and empathize with the

other's needs and vulnerabilities; and reluctance, based on stubborn pride, nursed grudges, and fear of humiliation, to initiate or respond to a positive, generous action so as to break out of the escalating vicious cycle in which they were trapped. (Deutsch, Coleman, and Marcus 2006, 1)

Intractable conflicts are in many ways similar. Just like tractable conflicts, they are caused by “moral and identity differences, high-stakes resources, or struggles for power and self-determination.” Intractable conflicts, further, have also serious consequences for those involved, as they are “often costly in human and economic terms, and can become pervasive, affecting even mundane aspects of disputant’s lives.” But in terms of what differentiates intractable conflicts from tractable ones, it is resistance to resolution that is the most salient for intractable conflict (Coleman 2006, 534). The usual example of conflict that is intractable, is that of Israel and Palestine. I shall speak more of it in a moment. Let me first compare intractable conflict with deep moral disagreement.

To repeat, moral disagreement is defined by two conditions: (i) involved parties disagree about moral principles or values and (ii) there exists no further way of settling the disagreement between moral principles or values. Does this definition overlap with that of intractable conflict?

First thing to notice is that not all features of intractable conflict are present in deep moral disagreement. For example, a disagreement may not have “high-stakes,” and it may not include “struggles for power,” but it can still be deep. Recall Parfit’s reference to Bernard Williams. Importantly, however, the most salient feature of intractable conflict – resistance to resolution – is indeed present in deep moral disagreement in the form of (ii). Further, a conflict’s being about “moral and identity differences” is a feature of deep moral disagreement too, as seen in condition (i). Definitions of the two notions, then, do not overlap, instead one includes the whole of the other. Let me give more concrete illustration.

For about last 80 years, there has been a conflict between Jews and Arabs in the land of historic Palestine. The conflict is very complex and disputed issues at its core have been changing throughout its history. At the moment, what is essentially at issue the most, is a question about whether there should be an internationally recognized Palestinian state next

to the state of Israel (Chomsky 2016). Analyzing this conflict, Donald Ellis (2020) identifies five characteristics that make it immune to resolution: it is (1) existential, meaning it is not about mere material resources, but about morality, human needs, or identity; it involves (2) power difference; (3) outgroup-bias; (4) “extreme emotions”; and (5) incommensurate descriptions of events or “narratives” (Ellis 2020, 184, for last characteristic see esp. 188-89). One way of understanding this list is to take it as a set of conditions for a type of intractable conflict to come about: if (1-5) hold, then conflict is irresolvable, and if it is irresolvable, it is intractable.

Notice that (1) can be easily substituted by (i) as both are essentially about morality. Further, if (i,2,3,4,5) hold, then there is nothing that can be done about this, or (ii). Therefore, if (i,2,3,4,5), then (ii). Deep moral disagreement is what partly constitutes the intractability of Israeli-Palestinian conflict. But this relation of inclusion may not always hold, as seen in the next illustration.

In contrast to the Israeli-Palestinian intractable conflict, take recently published case study of effectiveness of third-party mediation in resolving intractable conflicts (Boss et al. 2018). Here, the issue at question was a workplace disagreement in a hospital between a physician Mary and a surgeon Don, the result of which was “difficult working environment” (Boss et al. 2018, 243). Not a word is said about moral values or principles. All the reasons for the conflict were only what we may call pragmatic: Don did not like how Mary handles things and he wanted her replaced. This is why authors of the study omit any mention of morality in their definition of intractable conflicts, which according to the them are “prolonged disputes between two or more parties, which are resistant to constructive resolution efforts, destructive, and long-lasting” (Boss et al. 2018, 235). In terms of (i) and (ii), the intractable conflict between Don and Mary had very little to do with (i), but it still met (ii). What this means is that not all intractable conflicts may involve *moral* disagreements.

We can thus see that deep moral disagreement is sometimes constitutive part of intractable conflict, but not always. Or put differently, all the features of deep moral disagreement are found in *some* intractable conflicts. It follows from this that deep moral disagreements are part of what constitutes some intractable conflicts – Israeli-Palestinian conflict is but one example.

It is worth remarking that this conclusion is in line with conclusions by peace and conflict researchers (Mitchell 2014, chap. 11). In my view, it has been needlessly neglected by philosophical literature.

If I am right, then a new way of presenting historical examples of deep moral disagreements opens itself. To repeat, historical, in contrast to ahistorical examples of deep moral disagreements meet not only conditions (i) and (ii), but existence condition as well. That is, historical examples are not only cases of people being unable to resolve their disagreement on moral principles or values, but they are also *actually existing* cases. I believe many intractable conflicts to be these historical examples. There is a catch, however.

It is not like intractable conflicts are *impossible* to resolve, but rather that it is very difficult to do so. Thus, after 14 years of mediation and dialogue, Mary's and Don's disagreement mentioned above virtually disappeared (Boss et al. 2018). This was by no means an exception. There is other promising research showing successful attempts to mitigate intractable conflicts (Halperin and Pliskin 2015; Kapshuk and Shapira 2022).

What these studies imply, then, is the possibility, or hope, that (1-5) could be mitigated or eliminated. That is, after all, the main aim of conflict resolution strategies. However, does that in turn imply that the part of intractable conflict that is deep moral disagreement, i.e. (i,ii), disappears too? This, I believe, is not clear.

Surely a conflict can be resolved without people stopping deeply disagreeing on morality.⁶ This, for example, can be nicely seen in cases of compromise, another conflict resolution strategy. Compromise is “characterized by the fact that disagreeing parties hold on to their opposing views. [...] In a compromise, disagreeing parties *agree to* partially concede their claims to the demands of the other party, but they do not *agree with* the other party's demands”⁷

⁶ I wish to thank Kamila Pacovska for bringing this to my attention.

⁷ Compare this to consensus: “Unlike compromise, consensus requires the parties to a disagreement to change their minds on the controversial issue. If a consensus is achieved, this means that the disagreeing parties consider the agreement to be better than (or at least as good as) their initial positions” (Spang 2023, 2).

Take for instance South African conflict between white minority government, or National Party (NP), and African National Congress (ANC), which resulted in the end of apartheid in 1994. In early 90s, after many concessions by both sides, many of the conflict's features, such as power imbalances, out-group bias or extreme emotions, were mitigated or disappeared completely (Jolobe 2019). This cannot be said of deep moral disagreements between NP and ANC, however. As Zwelethu Jolobe puts it metaphorically in his recent book on the role of international mediation in ending the conflict: "there was no love lost between the [white minority] government and ANC" (Jolobe 2019, 1).

In what way, if any, did NP's and ANC's deep moral disagreement disappear? A proper examination of this question is beyond the scope of the present paper, I shall therefore give only a sketch of an answer I take to be probable.

When deep moral disagreement disappears, it can very well become what we might call *latent*. By being latent, the disagreement is not manifested, but it is disposed to be so. That is, we may not know the disagreement exist, until we bring up the controversial topic. Until we actually ask people what they think. Imagine here all the manifested disagreements when people from the whole of political spectrum get together for, say, Thanksgiving. In light of this, we must reconsider our understanding of (ii), or the fact that people can sometimes do nothing about their moral differences.

At the end of the day, when conflicts are ended, hands shaken and resolutions signed, it is very well possible for deep moral disagreement to not disappear entirely, but instead to take on a new form by becoming latent and moving to the background of everyday life. But surely if people *ignore* their differences, or "live and let live" so to say, then they do, in fact, change something about their disagreement. If this is the case, then the condition that people can do nothing about the disagreement they find themselves in, or (ii), must be interpreted in a different, weaker way. Here is one suggestion: there is nothing people can do about their moral disagreement, but that might one day change. I suspect that by seeing some intractable conflicts as instances of deep moral disagreements, we commit ourselves to this weaker interpretation of (ii). But be that as it may, these considerations in

no way affect the fact that we are now able to illustrate actually existing moral disagreements that are impossible to resolve.

5. Conclusion

In this article, I have argued that usually given examples of deep moral disagreements are never of actual events that happened and that this can, and should, be remedied by use of intractable conflicts. Deep moral disagreements are situations in which parties disagree on moral values or principles without having a way to settle the disagreement. On the other hand, intractable conflicts are situations of pervasive disagreements on existential matters that negatively affect involved parties on both emotional and physical level. Most importantly, intractable conflicts are notoriously difficult to resolve.

Deep moral disagreements are appealed to in different kinds of arguments. I mentioned two. The first was meta-ethical. It claimed that prevailing disagreements are best explained by denying the objectivity of moral values. The second illustration came from moral epistemology. It explored the thesis that the existence of deep moral disagreement undermines the possibility of “moral expertise.” Both of these arguments draw their strength partly on the fact that deep moral disagreements actually exist. I have been repeatedly showing that evidence, if we may take it as such, given in support of the existence of deep moral disagreement is weak, if not entirely lacking. This evidence takes the form of examples of generalized disagreements, such as debate on permissibility of abortion or morality of meat eating. These moral disagreements do, undoubtedly, occur – but it is not clear why we should, without further analysis of their particular instances, believe they are also deep.

It is sometimes argued that deep moral disagreement need not actually exist, because their conceptual possibility does the job just as well. This line of answer, however, presupposes that moral realism does not hold. That is, that moral claims are not propositions that one gets either right, or wrong. Without further argument against moral realism, then, this line of answer is not satisfactory. Therefore, there is still a good reason to try and find

actually existing deep moral disagreements. I suggested this can be done by looking at what peace and conflict studies call “intractable conflicts.”

Some of the features of intractable conflicts – namely them being about morality and difficult to resolve – are also features of deep moral disagreements. This means that when the former occurs, the latter also occurs. If I am right, some intractable conflicts are partly constituted by the fact that people disagree deeply on morality in them. I also mentioned that among social psychologists, this claim is uncontroversial.

Finally, my claim comes with theoretical baggage. If we grant that deep moral disagreements are constitutive part of some intractable conflicts, then we must amend our understanding of deep moral disagreement’s impossibility of resolution. This is so, because throughout history many intractable conflicts were resolved, if difficult. I suggested one interpretation, according to which deep moral disagreements can become hidden, or latent. This most often happens in cases of compromise. Here, it is obviously not the case that people can do nothing about their disagreement. I therefore suggested a weaker interpretation of this condition, according to which there is a possibility that one day, people might be able to do something about their deep moral differences – ignore them, for example.

Funding

The writing of this paper was supported by the "Internal grant of the Department of Philosophy and Religious Studies"

References

- Austin, J. L. 1979. “Ifs and Cans.” In *Philosophical Papers*, edited by J. O. Urmson and G. J. Warnock, 153–80. New York: Oxford University Press.
- Boss, Alan D., R. Wayne Boss, Benjamin B. Dunford, Matthew Perrigino, and David S. Boss. 2018. “Resolving Intractable Conflicts Through Third-Party Facilitation: A 14-Year Study.” *The Journal of Applied Behavioral Science* 54 (April): 234–71. <https://doi.org/10.1177/0021886318766014>.
- Coleman, Peter T. 2006. “Intractable Conflict.” In *The Handbook of Conflict Resolution: Theory and Practice*, edited by Morton Deutsch, Peter T. Coleman, and Eric Colton Marcus, 533–59. San Francisco, Calif: Jossey-Bass.

- Chomsky, Noam. 2016. *Fateful Triangle: The United States, Israel, and the Palestinians*.
- Deutsch, Morton, Peter T. Coleman, and Eric Colton Marcus, eds. 2006. *The Handbook of Conflict Resolution: Theory and Practice*. San Francisco, Calif: Jossey-Bass.
- Ellis, Donald G. 2020. "Talking to the Enemy: Difficult Conversations and Ethno-political Conflict." *Negotiation and Conflict Management Research* 13 (3). <https://doi.org/10.34891/1ftn-g083>.
- Fukuyama, Francis. 2006. *The End of History and the Last Man*. New York: Free Press.
- Gordon-Smith, Eleanor. 2019. *Stop Being Reasonable*. Sydney, NSW: New South Publishing.
- Haidt, Jonathan. 2013. *The Righteous Mind: Why Good People Are Divided By Politics and Religion*. New York: Vintage Books.
- Halperin, Eran, and Ruthie Pliskin. 2015. "Emotions and Emotion Regulation in Intractable Conflict: Studying Emotional Processes Within a Unique Context." *Political Psychology* 36 (S1): 119–50. <https://doi.org/10.1111/pops.12236>.
- Jolobe, Zwelethu. 2019. *International Mediation in the South African Transition: Brokering Power in Intractable Conflicts*. New York: Routledge Taylor & Francis Group.
- Kapshuk, Yoav, and Noa Shapira. 2022. "Learning about Dialogue and Partnership between Rival Groups during an Intractable Conflict." *Peace and Conflict: Journal of Peace Psychology*, July. <https://doi.org/10.1037/pac0000627>.
- Kekes, John. 1996. *The Morality of Pluralism*. Princeton, NJ: Princeton University Press.
- Mackie, John L. 1991. *Ethics: Inventing Right and Wrong*. Reprinted. Penguin Book Philosophy. London: Penguin Books.
- McGrath, Sarah. 2008. "Moral Disagreement and Moral Expertise." In *Oxford Studies in Metaethics 3*. Oxford, New York: Oxford University Press.
- Mitchell, Christopher. 2014. *The Nature of Intractable Conflict: Resolution in the Twenty-First Century*. London: Palgrave Macmillan UK. <https://doi.org/10.1057/9781137454157>.
- Parfit, Derek. 2011. *On What Matters*. The Berkeley Tanner Lectures. Oxford: Oxford University Press.
- Rowland, Richard. 2021. *Moral Disagreement*. New Problems of Philosophy. New York City: Routledge.
- Spang, Friderike. 2023. "Compromise in Political Theory." *Political Studies Review*, January, 14789299221131268. <https://doi.org/10.1177/14789299221131268>.

-
- Sturgeon, Nicholas L. 1994. "Moral Disagreement and Moral Relativism." *Social Philosophy and Policy* 11 (1): 80–115.
<https://doi.org/10.1017/S0265052500004301>.
- Tolhurst, William. 1987. "The Argument from Moral Disagreement." *Ethics* 97 (3): 610–21. <https://doi.org/10.1086/292869>.
- Thomson, Judith J. 1971. "A Defense of Abortion." *Philosophy & Public Affairs* (1) 1: 47-66.
- Williamson, Timothy. 2005. "Armchair Philosophy, Metaphysical Modality and Counterfactual Thinking." *Proceedings of the Aristotelian Society* 105 (1): 1–23. <https://doi.org/10.1111/j.0066-7373.2004.00100.x>.

On Everything Is Necessarily What It Is

Alex Blum*

Received: 26 June 2023 / Revised: 18 August 2023 / Accepted: 26 August 2023

Abstract: It is argued that if everything is necessarily what it is, then given the equivalence ‘ $p \equiv [a = (\forall x)(x = a \& p)]$ ’, it follows that whatever happens or is the case, had to happen or had to be the case.

Keywords: Fatalism; Identity; Necessity; The sole object; $(x) (\Box x = x)$

If we grant the equivalence

$$(1) p \equiv [a = (\forall x)(x = a \& p)],$$

that every sentence is equivalent to an identity sentence¹, and grant that (2) if a sentence is (necessarily) true then what it says is (necessarily) the case; then unless fatalism is true (3) the thesis of the necessity of identity,² is false, and thus so is the thesis that (4) everything is necessarily what it is³.


¹ Commonly assumed in one form or another by Church, Davidson, Gödel and Quine. See Yaroslav Shranko and Heinrich Wansing (2020). See Neale (2001: esp. 170-171).

² See, Kripke (1971,136).

³ The argument for the thesis of the necessity of identity rests on the formula ‘ $(x) (\Box x = x)$ ’. See Wiggins (1965:41) and Kripke (1971, 136). And is in fact equivalent

* Bar Ilan University

 <https://orcid.org/0000-0003-3912-1988>

 Department of Philosophy, Jakobovits Building, 1002, Bar Ilan University, Tel Aviv District City of Ramat Gan, Israel

 Alex.Blum@biu.ac.il

Suppose

$$(1) \quad p \equiv [a = (\exists x)(x=a \& p)]$$

is logically true, then given (2), so is,

$$(2') \quad p \text{ is the case if and only } a = (\exists x)(x = a \& p)$$

Hence given the necessity of identity, it follows that

$$(5) \quad p \text{ is the case if and only if necessarily } [a = (\exists x)(x = a \& p)]^4.$$

And thus,

$$(6) \quad \text{if } p \text{ is the case, then necessarily } [a = (\exists x)(x = a \& p)].$$

But if

$$(7) \quad \text{necessarily } [a = (\exists x)(x = a \& p)] \text{ then,}$$

$$(8) \quad \text{necessarily } p.$$

And thus,

$$(9) \quad \text{If } p \text{ then necessarily } p.$$

Hence, given that (1) is a logical truth and (2) is analytic, the thesis of the necessity of identity or the thesis that everything is necessarily what it is, implies fatalism.⁵

to it (Blum:x). We rendition the reflexivity of identity as ‘everything is what it is’. See Leibniz (1996, 362).

⁴ The argument for the necessity of identity is immune to whether the terms in an identity are expressed as ‘a’ or as ‘ $(\exists x)(x=a \& p)$ ’. Thus the argument will go through for:

$$a=(\exists x)(x=a \& p) \supset [Fa \supset F(\exists x)(x=a \& p)]. \text{ Let ‘F’= ‘}\Box\text{a’}.$$

We then have:

$$a=(\exists x)(x=a \& p) \supset [\Box a = a \supset \Box a = (\exists x)(x = a \& p)].$$

And thus:

$$a = (\exists x)(x = a \& p) \supset \Box a = (\exists x)(x = a \& p).$$

⁵ I am deeply grateful to Yehuda Gellman and to the reviewer for their comments.

References

- Blum, Alex (2023). "On the Argument for the Necessity of Identity". *Symposion*, Vol 10 Number 3 (forthcoming).
- Kripke, Saul (1971). "Identity and Necessity." In *Identity and Individuation*, edited by Milton K. Munitz, pp.161–91. New York: New York University Press.
- Leibniz, Gottfried Wilhelm (1996). *New Essays on Human Understanding: Second Edition*. Translated and Edited by Peter Remnant and Jonathan Bennett. Cambridge: Cambridge University Press.
- Neale, Stephen (2001). *Facing Facts*. Oxford: Oxford University Press.
- Shramko, Yaroslav and Heinrich Wansing, *The Slingshot Argument* in "Truth Values", *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2021/entries/truth-values/>
- Wiggins, David (1965) "Identity-Statements" in *Analytical Philosophy* (Second Series) edited by R. J. Butler. Basil Blackwell and Mott, Oxford UK, pp. 40–71.

Perfect Thinkers, Perfect Speakers and Internalism about Thought Content

Madelaine Angelova-Elchinova*

Received: 12 December 2022 / Revised: 15 August 2023 / Accepted: 12 September 2023

Abstract: In the following paper I propose an argument against internalism about thought content. After a brief preview of the recent debate between Sarah Sawyer and Åsa Wikforss, the paper outlines the central issue in their discussion. I show that, even if Wikforss' objections against Sawyer are granted, externalism about thought content can still prevail. For my argument, I use Wikforss' own objection against externalism and show how, if accepted, it binds one to the mythical figure of the perfect speaker – an infallible creature that possesses complete understanding of all of her concepts.


Keywords: Externalism; thought content; concept mastery; understanding; perfect speakers.

1. Introduction

Internalism about thought content (ITC) is the claim that the contents of our thought are internally individuated and not constituted by external factors. In this paper, I examine the debate between Sarah

* Sofia University

 <https://orcid.org/0000-0002-5961-0327>

 Department of Philosophy, Sofia University, 115 Tsar Osvoboditel Blvd, Sofia, Bulgaria

 m.angelova-elchinova@phls.uni-sofia.bg



Sawyer and Åsa Wikforss, aiming to show that even if all the arguments of an internalist like Wikforss are accepted and all her points made against Sara Sawyer are indeed granted, ITC is still a flawed thesis. Thus, my argument also presents an indirect defence of Externalism about Thought Content (ETC). However, instead of defending social externalism like the one championed by Tyler Burge (at least in his early works, for example Burge (1979)), I am taking a different approach by pronouncing myself in favour of Sarah Sawyer`s (2003, 2021) broad physical externalism. I am going to take the easy way by raising a rather modest, negative claim instead of a positive, more radical one – rather than arguing that ETC provides the only good explanation of concept possession or present arguments establishing ETC as the correct approach, I only demonstrate why internalism is unacceptable. The problem with ITC, as per my reconstruction of Wikforss` position, is that it begets the mythical figure of the perfect thinker (corresponding to a perfect speaker) - a creature that possesses infallible knowledge about how to employ the concepts it operates with¹. I hope to show that even if Wikforss` argument against Sawyer stands and externalism suggests that we possess incomplete understanding about our thoughts, that should not be a problem, because (usually) we have nothing in common with such creatures like the perfect speaker.

Section 1 serves as a preliminary section where I provide a reconstruction of Burge`s initial externalist claims and outline a clear-cut distinction between social externalism and broad physical externalism. I will start by making some introductory remarks about thought content and concept possession in general and then move to Burge`s Arthritis scenario.

In Section 2, I am going to provide an overview of the debate between Wikforss and Sawyer. I will be more concerned with granting Wikforss` arguments than with reviewing Sawyer`s replies as, while I am clearly an open sympathizer of Sawyer`s position, I still want to present the most

¹ If one is not too fond of the usage of the term ‘knowledge how’ here, it can be replaced by ‘some cognitive ability’ – that will not influence the central argument in any way.

charitable interpretation of the internalist's charges against broad physical externalism.

In Section 3, I will even grant that the main accusation made by Wikforss in regards to ETC is completely on-point - hidden in the articulation of ETC, there is indeed a tacit requirement to endorse the possibility of incomplete understanding of our own thoughts. Further, I show that the internalist alternative presumes that we adopt a Concept Mastery condition (CM) for concept possession. I examine different possible interpretations of CM and discuss possible ways to detect conceptual errors.

Finally, in Section 4 I build my argument by showing why it should be acceptable that one does not possess complete understanding of her own thoughts. The ITC alternative seems much more unpalatable, because by introducing CM, it also presupposes that we are perfect thinkers (corresponding to perfect speakers). I conclude by analysing the unreasonable implications of CM. I compare them to the implications of granting that partial understanding may be sufficient for attributions of concept-possession.

2. The content of a thought and the possession of a concept

It is unclear enough what our thoughts are supposed to be, but when it comes to their contents, things really start to look ambiguous. Tyler Burge makes some key points about the semantical foundations of the ITC vs ETC debate. The expression "thought content" can be regarded (non-strictly speaking) as synonymous to "mental content" or even to "conceptual content".

Before I introduce the basic definitions in the debate between holders of ITC and supporters of ETC, I would like to adopt two premises. First, I shall consider that thoughts have (at least some) sentential expression². It

² A legitimate worry about my presumption can be that it presupposes externalism from the start. One can argue, that if thoughts can (always) receive sentential expression it looks like they always conform to some social meaning. Davidson raises a similar objection against Burge, insisting that what we mean and think is not

would be hard to imagine that thoughts can be examined “on their own”, that is – without referring to any linguistic equivalents. For example, when dealing with the problem of conceptual truths, Williamson makes the following assumptions: a) concepts are the constitutive elements of thoughts, b) “to grasp a thought is to entertain it” (Williamson 2006, 2) and c) thoughts are expressed via sentences (ibid, 9: footnote 9). Thus, in Williamson’s case, the analysis of conceptual truths is done as an analysis of their sentential expressions as public linguistic elements. Following a similar approach, the second premise that I adopt can be formulated as follows: Public access to the content of a thought is provided (in principle) by its sentential expression³. I suppose that both premises are pretty straightforward and their admittance should not present a problem⁴.

A short reconstruction of Burge’s key points made in relation to the nature of thought content is also in order, as the definitions provided by him can be regarded as widely accepted and, more importantly for the purpose of this paper, they are the same definitions as the ones used by Wikforss and Sawyer.

Thought contents can be characterised as obliquely occurrent expressions in sentential that clauses (or content clauses) (Burge 1979, 76), e.g. Maddy believes *that* Earth is round. As Burge himself points out, the

necessarily determined by the linguistic habits of those around us (1987, 448). Thus, thoughts can still have “narrow contents”, which are internally individuated. I would like to thank Johan Heemskerk for pointing this out to me. The problem is that such internalism will face namely the dilemma Wikforss wants to avoid: giving up on reference or rejecting the communitarian conclusions, thus accepting conceptual and referential fragmentation (for Wikforss’ own worries about the internalist dilemma see 2001, 217-218; for her worries about Davidson’s answer to Burge see 2001, 227).

³ The first premise can be regarded as more general because it suggests which background theory of thinking I am adopting. Whilst the second premise is implied by the first, I still consider it different as it informs the reader about the particular way in which I am employing the expression “thought content”.

⁴ The adoption of such premises can seem to imply that conceptual mastery is related to linguistic mastery. Unlike the premises themselves, this conclusion may give rise to a very strong objection. In Section 4, I address some worries that may spur from the relation between concept possession and linguistic expression.

terminology employed by him is one that he borrows from “the mentalistic discourse” (“the intentional discourse”). Thus, thought contents are presupposed to possess representational (intentional) character and to reflect one’s epistemic perspective⁵ (Ibid.). More importantly, it is presumed that in cases where extensional differences between two obliquely occurring counterpart expressions in that clauses are presented, we can examine them as describing two different mental states or events (Burge 1979, 77). The tacit presumption (as it will become clear from the Sawyer/ Wikforss debate) is that the difference between two mental states originates from a difference in their reference. Finally, as it is broadly accepted that the content of a thought can be regarded as constituted by concepts⁶, the remaining pages are going to deal directly with concept possession and conceptual content.

Now, moving to the ITC vs ETC debate, a well-articulated general suggestion on how to distinguish internalists from externalists with regards to a property K, is laid out by Mark Rowlands, Joe Lau and Max Deutsch:

In its most general formulation, *externalism* with regard to a property K is a thesis about how K is individuated. It says that whether a creature has K or not depends in part on facts about how the creature is related to its external environment. (...)

Individualism or internalism with respect to a property K says that whether a creature has K or not supervenes on its intrinsic properties only. It follows that facts about the environment play no role in determining whether or not the creature has property K. (Rowlands, Lau and Deutsch 2020).

In his Arthritis case, Burge uses the same criterion to show that thought content is not completely determined by one’s intrinsic properties. He

⁵ The presumption that there is an intimate relation between thought contents and epistemic perspective provides yet another reason why we must take heed of the problems surrounding individualism as they have important implications for general debates in epistemology (e.g. issues about the nature and possibility of a priori knowledge).

⁶ Burge prefers to talk about “notions”, because this term is more isolated from theoretical commitments. In my paper, I go with the more traditional one - “concepts”.

invites us to imagine a scenario with two individuals – let’s call them Bert and his Counterpart – that are identical in every important aspect which concerns their ‘inner life’ (viz. mental states, desires, propositional attitudes) as well as their hard-wired neurological setting and all of their neural processes (Burge 1979, 77-78). Burge also makes the important clarification that both Bert’s and Counterpart’s disposition to assent to the proposition $p = I \text{ have arthritis in the thigh}$ is caused by the same processes and can be traced to the same factors (ibid, 78). By introducing a difference in their external, social environment, Burge suggests that even if they are intrinsically indistinguishable, they possess different thought content. While Bert’s proposition p is false, his identical counterpart’s proposition p is, nevertheless, true⁷. The only difference between Bert and Counterpart lies in their social environment. Bert shares our social environment where ‘arthritis’ does not apply to ailments outside the joints. As a result, his belief that p is false. In Counterpart’s environment, “physicians, lexicographers, and informed laymen apply ‘arthritis’ not only to arthritis but to various other rheumatoid ailments” (Burge 1979, 78). As a result, Counterpart’s belief that p is true.

The view defended by Burge in “Individualism and the mental” (1979) is usually called ‘social externalism’ and it can be considered pivotal for ETC theories. However, a more recent defence of ETC, which is properly constructed as ‘broad physical externalism’ is presented by Sarah Sawyer. The difference between these two variations of ETC can be summarized as follows:

Social Externalism (SE) [Burge in IM (1979)]: The individuation of mental states or events and the forming of corresponding propositional attitudes is dependent on one’s social environment (Burge 1979, 84-85; Sawyer 1993, 265; Wikforss 2001, 217).

Broad Physical Externalism (BPE)⁸ [Burge in his later works and Sawyer]: Two individuals A and B can be physically identical while having different mental states and this difference is not instantiated by

⁷ Propositions expressing thought content always possess fixed truth values, e.g. two identical propositions with different truth values also express different thought contents.

⁸ Sawyer also refers to this position as “natural kind externalism” (Sawyer 2015).

anything in their ‘mental’ life but is dependent on their environment, which is broadly construed as a physical environment (Burge 1986, 708; Sawyer 2003, 272).

There can be important differences with regards to the possible implications of SE and BPE, which I will not present in detail here. As far as the argument that I propose is one favouring Sawyer’s claims in the debate against Wikforss, maybe it is more precise to say that it is an argument that favours BPE. However, I do not see any obvious reason why the argument presented in Section 3 should not be applicable to versions of SE.

What further divides ETC from ITC⁹ are views about what constitutes concept possession. All participants in the debate seem to agree that in order for S to entertain a thought T, S has to possess understanding of the constituting (in regards to T) concept C. What they do not agree upon is whether such understanding should be immaculate. I am going to use a distinction between degrees of understanding introduced by Gabriel Rabin (2020) to illustrate the disagreement between ITC and ETC supporters. According to Rabin, *concept possession* is what allows an individual to entertain a particular thought containing this concept (Rabin 2020, 627). For example, to be able to think that one has hands, S has to possess the concept C = HANDS. However, if S grasps and applies C correctly on any occasion, then we can say that S possesses full understanding of C. In such cases we can ascribe *concept mastery* of C to S (ibid, 627-628). Going further, concept possession can either coincide with concept mastery or it can allow for partial or incomplete understanding.

The question whether concept mastery is necessary for concept possession has become an apple of discord between holders of ITC and ETC. While externalists insist that a subject S’s partial understanding (incomplete mastery) of a concept C can and, on many occasions, does present a sufficient condition to ascribe concept possession of C to S (Burge 1979: 83-84), internalists contend that only concept mastery will suffice¹⁰. One way

⁹ The disagreement in question would be especially applicable to the debate between Wikforss and Sawyer.

¹⁰ Sarah Sawyer (2003) actually argues that SE does not need to rest on incomplete understanding or the possibility of a conceptual error. To strengthen the position of

to go about settling issues regarding the nature of thought-content is to focus on our ability to apply and grasp concepts. Thus, the cornerstone of my argument will be a) to examine what exactly is required for concept possession and b) to settle on the most viable interpretation of the concept mastery condition and show its implausibility.

3. Incomplete understanding

In the previous section, I argued that in order to pick a side in the debate between ITC and ETC one should gain awareness of rival interpretations of the requirement for concept possession. In this section, I provide an overview of the debate between Åsa Wikforss and Sarah Sawyer. As I read it, the bedrock of their discussion is exactly their disagreement about what constitutes concept possession. A careful look into Wikforss' objections against ETC allows one to infer the following central accusation: The defence of SE (and BPE by extension) depends on i) the condition that in Burge's example, Bert/Counterpart makes a conceptual error and ii) on the acceptance of the notion of incomplete understanding (Wikforss 2001, 2004). Wikforss insinuates that it is strange to suppose that we do not understand our own thoughts. Partial understanding looks like a necessary presupposition in regards to ETC (Wikforss 2004, 287). In Burge's wider (i.e. broad physical) externalism, the "incompleteness" becomes even more "radical and pervasive"¹¹ (Wikforss 2004, 294).

the internalist, I will presume that her argument fails and that externalism does require cases of partial understanding to count as cases of concept possession.

¹¹ The idea that full understanding of our thoughts seems natural and intuitive spurs from the same assumption that Williamson discusses in "Cognitive homelessness" – namely, that nothing in our mind nor in "the realm of philosophy" remains hidden from us (Williamson 1996, 554). It is quite natural that ITC supporters are also inclined towards accepting Descartes's presumptions about our "cognitive mobile homes" without hesitation. However, the readily adopted claim that all of us have such cognitive structures which remain open to us at all times turns out to be quite problematic (ibid) (for a rebuttal of Williamson's anti-luminosity see Berker 2008).

Wikforss does not give any particular reason in support of her claim that the notion of incomplete understanding is unfeasible, nor does she provide any justification to support the idea that concept mastery is required for conceptual understanding. It is left to the reader to reach to the same conclusions by relying on her intuition or common sense. And while it may be true that on account of our *prima facie* intuitions incomplete understanding of our own thoughts sounds unreasonable, I am going to show that the alternative is far more unsatisfactory. One more thing that is somehow left to the reader is how to interpret the concept mastery (full understanding) condition. While all parties in the debate provide some insight into what is involved in partial/incomplete understanding and why it should/should not be regarded as sufficient for concept possession, nothing much is said regarding the more rigid requirement. That is why a careful examination of possible interpretations of the concept mastery condition will follow in Section 3.

Before engaging in clarification of the notion of concept mastery, let's first examine some of the key points in the debate made by Wikforss and Sawyer. In regards to Wikforss's objection that ETC invokes the concept of incomplete understanding, the answer provided by Sawyer which looks most promising is that "the unifying principle of externalism" does not require that a subject S would still be able to grasp a concept C in cases where S has only incomplete understanding about C. A characteristic of all versions of ETC is rather the acceptance of the claim that concepts are not only individuated by a subject's psychological states, e.g. by what the subject thinks is true of a given referent, but, also, by the referent itself¹² (Sawyer 2003: 272). Such principle can be regarded as "a unifying principle of reference" (UPR) and reformulated as follows: "*a difference in reference (extension) implies a difference in concepts*" (Wikforss 2004: 290).

Turning to Sawyer's appeal to the unifying principle of reference, Wikforss raises the objection that depending on how we construe UPR, it is either false or begging the central question (Ibid. 291). She goes on to show,

¹² Sawyer's answer provides a straightforward explanation of why such theories are 'externalist'. Referents are understood not as determined by one's individual psychological processes but as elements of some independent reality.

using Burge's Sofa case¹³, that on the first interpretation of UPR a term can have different extension without that implying a difference in concepts. For example, on this interpretation the extension of 'sofa' can be different (we can imagine a possible world where all sofas are made of leather) while the meaning of 'sofa' remains exactly the same.

On the other hand, on the second interpretation of UPR, our term 'sofa' would not apply to objects in the counterfactual situation, because they will not fall into the extension of 'sofa' in our world (i.e. they will not *be sofas*). Interpreted like that, UPR is true but it commits one to a type of externalism presupposing that incomplete understanding is sufficient for concept possession, namely – reference externalism. It is crucial to see precisely how Wikforss' objection is spelled out, as it will prove central for my own argument:

“Construed in the first way, the principle [M.A. UPR] is false, since there are many possible worlds in which our term 'sofa' has a different extension without thereby expressing a different concept. Construed in the second way, the principle is true, and can be used to defend the move from reference externalism to content externalism, but not to support reference externalism in the first place. To make a case for the claim that our term 'sofa' does not apply to the objects in B's world, considerations of a quite different kind are required. The question, then, is **whether these considerations will commit the externalist to the assumption of incomplete understanding**¹⁴.” (Wikforss 2004, 292)

The point made by Wikforss is that even if UPR can indeed justify the move from reference to content externalism (i.e. to ETC), the grounding of

¹³ The Sofa case aims again to establish that thought contents cannot be individuated simply qua psychological (internal) processes by arguing that two physically indistinguishable subjects – A and B - can have different mental contents and that therefore mental content is externally individuated (for an excellent reconstruction of the Sofa case see Sawyer 2003: 267-268).

¹⁴ I have intentionally put the last part of the objection in bolded text and I would urge the reader to keep in mind that Wikforss considers incomplete understanding as the sole failure of ETC.

ETC on reference externalism is exactly what requires the defender of ETC to accept the assumption of incomplete understanding¹⁵. For the sake of simplicity, let's call this argument The Argument from Incomplete Understanding (AIU). There is a way to rephrase AIU to make its elements clearer:

AIU

- P1: The unifying principle of reference can be used to justify the transition from externalism in regards of reference to ETC.
- P2: Arguments in favour of externalism in regards of reference often rely on cases where a subject S possesses only incomplete understanding about a concept C.
- C1: ETC would also require that we allow for concept possession in cases of incomplete understanding.
- P3: It is unacceptable (or a demonstration of “radical and pervasive” incompleteness¹⁶) to suggest that we do not understand our own thoughts.
- C2: The main claim of ETC is false.

The approach preferred by Sawyer (2018) is seemingly to reject C1 by showing that it fails to account for the fact that there are nonrepresentational relations among the content of S's thoughts and some objective properties in her wider physical reality¹⁷ (Sawyer 2018, 5). I, on the other hand would like to propose a different line of defence for ETC. I intend to show that even if premises P1 and P2 are to be accepted and even if the conclusion

¹⁵ When Sawyer argues that the unifying principle of externalism is the principle of reference, she somehow tries to detach her views from those explicitly expressed by Burge, who insists that we can ascribe to S a possession of a concept C even in cases where S has only partial or incomplete understanding about C (Burge 1979, 79).

¹⁶ I appeal again to the exact wording which is used by Wikforss (2004, 294).

¹⁷ Another example that suggests Sawyer's withdrawal from C1 can be found in her introductory paper on “Internalism and Externalism in Mind”. There, she draws an explicit distinction between Burge's early SE and the one endorsed by her during the debate with Wikforss (BPE) based on the fact that only SE obliges one to accept incomplete understanding as plausible (Sawyer 2015, 135-137).

C1 in AIU is indeed true, the conclusion C2 would still be false. I consider P3 to be the weak link in AIU. It turns out that even if we imagine that Wikforss is right about almost everything and Sawyer is wrong¹⁸, even if externalism does depend on the concept of incomplete understanding, one can simply argue that this is not a problem.

Before arguing in favour of the possibility of incomplete understanding, it is crucial to set the stakes more clearly by examining how the notions of concept mastery and complete understanding can and should be interpreted.

4. Concept mastery

I have argued so far that in order to decide if ITC or ETC fares better in providing an explanation of content-individuation, we should determine what is required for concept possession. I submit that a rejoinder to Wikforss' main challenge to ETC can use her own argument against her by showing that incomplete understanding of our own thoughts and conceptual error are indeed possible. By accepting C1 of Wikforss' argument, the only thing that one has to do in order to salvage externalism is to show that P3 is false.

A careful look into Wikforss' argument allows one to see that if the claim expressed by P3 turns out to be false, then the first portion of AIU can serve a reverse purpose as an argument in favour of ETC. Let me explain what I mean by that. If it turns out that there is not a general problem with the ascription of concept possession to individuals who do not completely understand a concept, then (by Wikforss' own admission) incomplete understanding can be used to "support reference externalism in the first place" (Wikforss 2004, 292). However, as Wikforss admits, if reference externalism is supported by something different than the unifying principle, then the principle can be true (as insinuated by P1), and can "be used to defend the move from reference externalism to content externalism" (ibid).

¹⁸ I do not consider Sawyer's claims misguided, on the contrary – I am a great sympathizer of her ideas. However, I think that there is a better approach to Wikforss' challenge.

Therefore, a reasonable rejection of P3 would not only undermine individualism but, also, it will turn AIU into a strong defence of ETC.

In her effort to show that incomplete understanding is not enough for concept possession (which is at the heart of P3), Wikforss implicitly adopts the opposite premise which will be designated as ‘The Concept Mastery’ (CM) condition:

CM: We can ascribe a possession of a concept C to S if and only if S has full understanding about C (i.e. only in cases where concept mastery is presented).

Apart from appealing to a strong intuition, P3 does little work in providing a legitimate worry against ETC. However, when it comes to a philosophical dilemma, a strong intuition cannot just be ignored. Therefore, in the remaining pages of this paper I will try to show why P3 may seem intuitive, but also why it is wrong. To do that, first I have to find the most viable interpretation of the CM that gives the best possible chances for Wikforss’ argument to stand.

A difficulty for interpreting concept mastery is that while Burge, Wikforss and Sawyer do provide useful insight into what they mean by ‘partial’ or ‘incomplete’ understanding, very few remarks have been made when it comes to full understanding or concept mastery. Therefore, the construal of a good interpretation of CM will have to make use of the notion of incomplete understanding in order to infer what full understanding stands for.

A second difficulty arises in regards to how to detect concept mastery or concept possession or, in other words, which linguistic expressions of content-states signal appropriate concept possession¹⁹. On numerous occasions, Burge suggests that subjects in the Arthritis case and the Sofa case have “incomplete linguistic understanding” or “incomplete mastery of terms” but nonetheless can be regarded as possessing the relevant concepts (e.g. Burge 1979, 80; 1986, 708). He even suggests that a good portion of our beliefs (e.g. beliefs about what beef brisket is) are “infected by

¹⁹ As it will transpire in Section 4, another important question will be which linguistic expressions indicate conceptual error and which correspond merely to a linguistic error.

incomplete understanding”, but that does not necessarily imply that we lack conceptual competence (Burge 1979, 79-80).

The parallels between incomplete linguistic understanding and lack of concept mastery suggest one possible way to interpret the notion of concept mastery:

Accepted Usage construal (CM): S possesses a concept C if and only if S uses the linguistic expression L corresponding to C correctly or in accordance with the conventional meaning of L on any occasion²⁰.

Unfortunately, the Accepted Usage construal of CM is extremely implausible, not only because a) it gives rise to a Strawman argument against ITC by representing individualism as a very easy target, but also b) because of a discrepancy with some of Burge’s central claims. In the same paper where Burge seems to encourage such a construal (see Burge 1986), he also draws an important distinction between cognitive value (potential information units) and conventional meaning corresponding to “the gap between accepted usage and belief” (Ibid, 214). Thus, subjects satisfying the CM condition cannot just be regarded as impeccable players in a language game²¹.

Another way to interpret concept mastery is suggested by Burge’s criticism of the Cartesian interpretation of privileged access in “Individualism and the mental”. When arguing that authority of one’s reports about their thought contents applies to cases of incomplete understanding, he suggests that ITC presupposes some “special intellectual vision of the contents of (...) thoughts and beliefs” (1979, 116). Such an interpretation is further supported by Wikforss’s reading of conceptual error as a failure of the individual to grasp conceptual connections (2004, 294). CM forbids conceptual errors which allows us to conjecture that not only conceptual contents, but also conceptual connections should be *transparent*.

²⁰ Some passages in Burge’s work seem to suggest a similar construal (see Wikforss 2001, 224). On one of the very few occasions where he examines concept mastery (that is “full understanding of cognitive value” on his terminology), he describes it as “normally not distinct from ideal understanding of ordinary usage and meaning” (Burge 1986, 718).

²¹ Wikforss’s examination of Burge’s Sofa case provides a detailed argument prohibiting the Accepted Usage construal (Wikforss 2004: 293-294)

On this interpretation, CM presents a very rigid restriction – full understanding of the conceptual contents constituting a thought *T* is a precondition for *thinking* and rationality²²:

Strong (Positive) Interpretation (CM) – S possesses a concept C if and only if the content of C and all conceptual relations in which C is a relatum/relata are luminous for S²³.

But would the internalist, particularly Wikforss, accept such a strong construal? It is immediately evident that understood in this way CM is regressive and that full understanding would require thoughts about thoughts²⁴. Further, the strong interpretation is open to various lines of argument, most notably Burge's own remarks about unconscious possession of concepts (1979, 105) and variations of Williamson's (1996, 2000) anti-luminosity argument. More importantly, Wikforss never explicitly appeals to privileged access or mentions *awareness* of conceptual contents. Therefore, we can conclude that the Strong reading of CM is too uncharitable.

One remaining option is to treat the CM condition as a negative requirement. As Wikforss insist on multiple occasions, AIU stresses that ascriptions of concept possession presuppose that we are *not* in a possession of our own, deviant concepts (2001, 231), we are *not* failing at grasping important conceptual connections and we are *not* "rejecting trivial analyticities"²⁵:

²² cf. "Failure to grasp one's mental contents results from either blind prejudice or interference by "mere" bodily sensations and corporeal imagery." (Burge 1979: 104)

²³ Berker's interpretation of Williamson's anti-luminosity argument clearly shows that luminous conditions refer to "a kind of epistemic privileged access"

²⁴ I would like to thank an anonymous reviewer who suggested to me that this interpretation is too strong and uncharitable.

²⁵ Wikforss seems to suggest that cases of conceptual error are either "radical and deep-running" disagreements in meaning, but not in beliefs and rejections of trivial analyticities (2001, 231). Characterised like that however, CM is begging the question, because it leaves very little space for the possibility of conceptual error (e.g. trivial cases like "brisket" or radical examples like the Sofa case are neither disagreements in meaning nor they concern a trivial analyticity).

Weak (Negative) Interpretation (CM) – S possess a concept C if and only if, when it comes to C, S is a) unable to commit conceptual error and b) C is not a deviant concept.

Weak CM still involves one rigid restriction: if we grant that S possesses C, that amounts to granting that it is impossible that S ever makes a conceptual error in regards to C. In order to avoid circularity, we need a clear answer of what constitutes a conceptual error and how to detect it.

First, a conceptual error suggests partial grasp of a concept. However, we are looking for construal of such partial grasp in terms different from the ones provided by the Accepted-usage and the Strong interpretation. Further, we want partial *grasp* to reflect incomplete *understanding*²⁶. The proper way to go, then, is to present conceptual competence in terms of ability. As suggested by Sawyer, “the subject’s grasp of a concept is tied to the ability to apply this concept correctly” (2003, 271). Here, “correctly” designates “in terms of reference” (e.g. Burge 1986, 715, Wikforss 2001, 23). Thus, incomplete understanding/partial grasp presupposes “an ability to discriminate some but not all Fs from non-Fs typically” (Sawyer 2003, 271).

The aforementioned clarifications allow for a final, positive twist on the Weak interpretation of the CM condition:

Ability Interpretation (CM): S possess a concept C if and only if S possesses infallible knowledge how (or some other cognitive ability) to employ C correctly²⁷.

It is important to stress out that on the Ability interpretation, CM does not involve a Luminosity condition – the contents of one’s thoughts and concepts need not be transparent and readily accessible (at least not in a reflective manner). One satisfies the condition if one employs their concepts

²⁶ The reason for that is quite obvious. While Wikforss holds that “a linguistic mistake is not a conceptual mistake” (2001, 230), she insists on concepts being related to meaning as public and shared (c.f. her argument against Davidson in Wikforss 2001: 227)

²⁷ Note that S does not need to be aware of this ability and can even consider they lack such ability. At the end of the day, if CM is true, then it should also apply to people like Burge and Sawyer who do not find it “transparent”.

correctly and avoids conceptual errors, but one may yet be unable to determine when they satisfy the CM condition.

In the next section, I am going to argue that – put this way – CM is false. Now that the question about how to interpret concept mastery is settled, only one important clarification remains: we still have to precisify how to detect cases of conceptual error or how to determine if one possess such infallible knowledge how to employ a concept. I propose that we think of full understanding as a *discrimination* ability in terms of reference (per Sawyer’s proposal). Such interpretation is further plausible, because it is permitted (and even implied) by P2 in AIU. If Wikforss’ suggestion is that externalism in regards of reference requires incomplete understanding to be a viable option, then on her own, internalist account, the CM condition should also regard meaning in terms of reference. After all, Wikforss’ own claim is that her argument from incomplete understanding can vindicate internalism “without having to accept conceptual and referential fragmentation”²⁸ (2001, 218).

5. Perfect thinkers and perfect speakers

Now, I turn to the question: what if externalists can ‘bite the bullet’ and show that the commitment to incomplete understanding is a reasonable price to pay. I remind the reader yet again that in P1 of AIU Wikforss stipulates that if reference externalism is supported by something different than the unifying principle, then the principle can be true and can ground the move from reference externalism to ETC. Thus, if the CM condition embedded in P3 turns out to be false, AIU is actually giving us good reasons to endorse ETC.

In my argument against AIU, I use *reductio ad absurdum* stipulating that it would suffice to show how the adoption of the CM condition comes with unacceptable implications²⁹. I am going to do this by suggesting that

²⁸ If the argument was not aimed at rejecting the dilemma in front of the internalist, it would not have drawn attention to begin with.

²⁹ It is important to note that the CM requirement is not merely an idealisation which is supposed to show how real conceptual possession should look. CM is not a

the notion of concept mastery begets the mythical figure of the perfect speaker (as an extension of the perfect thinker). To allow that there are people or, which seems even more radical, that all people are such as to satisfy the CM condition, means to allow that they are creatures who possess infallible knowledge about how to employ the concepts they operate with. However, each concept they operate with is part of their thought content, which in turn is expressed via language. Therefore, our ‘infallible connoisseurs’ would also be perfect speakers due to their ability to use a linguistic expression L correctly on any occasion on account of their full understanding of a concept C which corresponds to said L³⁰. In other words, if we can ascribe concept mastery to S in regards to, say, the concept MEAT, that means a) that S is incapable of making a conceptual error when it comes to MEAT and b) that under normal conditions³¹ S is going to use the corresponding linguistic expression “meat” correctly (in terms of reference) in every sentence uttered by her.

One may object that such conclusion oversimplifies the matter and that making a linguistic error does not amount to making a conceptual error, nor does concept mastery presupposes linguistic mastery.³² This objection deserves attention and calls for some additional argument in favour of the relation between concept possession and sentential expression of concepts. First, I admit that a linguistic error on its own does not guarantee that one

normative condition in the sense that it does not just set the bar for how concept possession should be, but it rather suggests that concept possession really works like that.

³⁰ Let’s take the concept PLATIPUS. The Ability interpretation of CM is compatible with S having false beliefs containing the concept PLATIPUS, e.g. S may falsely believe that there is an angry platypus under the bed. What Ability-CM is incompatible with, is S making false judgments about the concept PLATIPUS and its referent – the natural kind platypus, say that platypuses do not produce venom. Wikforss’s argument allows this restriction, because she sustains that conceptual disagreements are disagreements where we share a lot of common beliefs about e.g. arthritis, but we fail to converge upon some of our beliefs about ARTRITIS (Wikforss 2001, 231).

³¹ “Under normal conditions” is meant to exclude cases of purely linguistic error. In what follows I will provide further clarifications.

³² I would like to thank an anonymous reviewer who pointed this out to me.

has not mastered the concept in question. There can be many explanations of why someone makes a linguistic error, for example the infamous lapsus linguae. Purely linguistic errors can also include cases of false beliefs about a term's application, e.g. Burge's subject who believed he had orangutans for breakfast (Burge 1979, 90-91; Wikforss 2001, 231). Particular types of linguistic error may also be due to a serious condition like dyslexia. Recent findings report dyslexia to be primarily related to a word identification problem due to issues with phonological processing, which nevertheless do not to presuppose comprehension failure (e.g. Casalis 2004).

Thus, there are indeed cases where linguistic error is not related to conceptual error. On the other hand, there are cases where conceptual error is made, but no (literally understood) linguistic error is presented. Let's look at an example:

Dolphins: Let's imagine that 10-year-old Martha loves dolphins. She goes to the Dolphinarium regularly where she observes the habits of the dolphins, swims with them regularly and feeds them fish. Martha has a variety of true beliefs with content-clauses involving oblique occurrences of DOLPHIN. For example: that dolphins are highly intelligent, that dolphins can swim, that dolphins eat fish and that dolphins produce a variety of vocalizations. Let's further imagine that Martha engages in a discussion with Peter who asks her which is her favourite fish. Martha answers the following:

S1: Dolphins are my favourite fish.

Are we to attribute possession of the concept DOLPHIN to Martha? I would say that we are. Now, according to the CM condition this would be a clear example of conceptual error and it would suggest that Martha has only partial understanding. Furthermore, there is no obvious linguistic error in S1.

One can wonder, does not this make the case against my assumption that concept possession is related to linguistic expression even stronger? My answer will be 'No, because that is a Strawman type of argument'. I insisted that CM implies that one has *to use* the corresponding linguistic expression e.g., "dolphin", *correctly* (in terms of reference) in every sentence uttered by her. That does not mean that she cannot make a *purely* linguistic error

in the sense of lapsus linguae or even in the sense of dyslexia, nor in any sense that concerns language rules or even accepted usage of expressions alone. Martha's error is one that concerns the meaning of "dolphin" in terms of reference, her belief that dolphins are fish is a mistaken belief *about* the concept DOLPHIN³³.

Thus, all that will be presumed by my perfect speakers charge is that such creatures have to be infallible in regards to *linguistic meaning*³⁴ (understood in the abovementioned way). All of the abovementioned examples show only that there is no necessary relation between concept possession and (literally understood) sentential expression. What they failed to show is that there is no relation between concept possession and sentential expression (understood in terms of meaning) and, also, that there is any other way to assess concept possession other than analysis of sentential expressions. If conceptual error was never presented in sentential expressions, there would be no way to pick it up. However, such presentation does not presuppose any *purely* linguistic error, only a reference error.

An internalist (mind you, one who's views are much closer to those of Davidson rather than those of Wikforss) may stipulate that Martha has a DOLPHIN-like concept, and she was correct with respect to that concept: her concept, whatever it included, was consistent with dolphins being fish³⁵. On this view, concepts are literally *individuated* and pertain solely to a given individual³⁶. Thus, Martha ends up in possession of the concept

³³ One can object that Martha's error is rather a factual error. I do not deny that. Even so, it is also a conceptual error, at least per everything that we can gather from Wikforss's interpretation of conceptual error. Martha does not have a deviant DOLPHIN concept, nor does she make a purely linguistic error.

³⁴ That is, for the corresponding linguistic expressions of concepts they are in possession of. Imperfectness would be possible, but it would suggest that in all such cases CM does not hold and that possession of the relevant concept cannot be attributed to S.

³⁵ I would like to thank Johan Heemskerk for suggesting this line of defence on behalf of the internalist.

³⁶ The scope of this paper does not allow that I dive into the metaphysics of concept in detail. My initial response to a naturalistic charge presuming that concepts are not abstract entities is that I agree with it. I even consider externalism to be far

FOLPHIN and we can grant that she mastered the concept FOLPHIN. I share Wikforss' scepticism that this line of argument does not do much good to the internalist. To paraphrase: If whenever two speakers disagree about the classification of dolphins it follows that they must have 'different thoughts', then it is hard to see how two speakers could ever share any thoughts at all³⁷. That would preclude us to share any concepts or disagree over our beliefs (Wikforss 2001, 227).

It is an open question if one is ready to attribute concept possession to Martha in the Dolphin case. While I am ready to do so, many would find that Ability-CM holds for Dolphin and would deny that Martha possesses the relevant concept. However, Dolphin is not meant to present a direct challenge of Ability-CM. The example's main purpose was to illustrate a common case of conceptual error. Martha is purposefully depicted as a child who, even if indeed infatuated with dolphins, is not an expert in any way and does not employ her DOLPHIN concept flawlessly. But is it possible that a lot of us are much more like Martha than we are ready to admit?

Going further with the argument against the CM condition, the first problem that it faces is that it is too rigid. We should allow that there are competent subjects who are capable of conceptual errors. Let's take the following example:

Substitute: Let's assume that Mike is a chef and works at a restaurant that mostly serves grill and barbeque. He possesses a large number of beliefs, which are commonly attributed with content clauses containing 'grilled chicken' in oblique occurrence. At this stage we can probably say that, as an expert and reliable user of the expression 'grilled chicken', Mike also seems to possess the concept GRILLED CHICKEN which corresponds to 'grilled chicken'. However, imagine that Mike has a friend

more capable to incorporate a naturalistic view on concepts than internalism, e.g. Sawyer's natural-kind externalism (2015).

³⁷ Further, another worry is that DOLPHIN-like concepts are a slippery slope, until you realize it, you have a million different concepts for a dolphin and none of them is DOLPHIN, because there is no concept DOLPHIN anymore. Getting rid of reference may fend off the problems raised by Putnam, Burge and etc., but what does it leave us with? That is why I agree with Wikforss that a Davidsonian solution is a non-starter.

named Judith. One night, Mike goes to diner in Judith`s house. Unbeknownst to Mike, Judith has prepared a dish containing only soya chunks and vegetables. While eating his dish, Mike utters the sentence *s1*: “I think that this is the best grilled chicken I ever tasted”. Should we suspend our initial judgment and deny Mike the possession of the concept GRILLED CHICKEN?

I contend that examples like Substitute are a useful demonstration of why we should restrain from a foolhardy acceptance of the CM condition. Clearly there are instances in which we would usually ascribe concept possession (i.e. we would also suppose that partial understanding is sufficient) even if, due to unforeseen circumstances, upon occasion S makes a conceptual error. Just like Bert in Burge`s example, Mike makes such a conceptual error expressed in *s1* by forming a false belief *about* grilled chicken.

But why should we presuppose that Mike is making a conceptual and not simply an empirical error? Remember that Ability-CM presupposes that we should interpret full understanding as a discrimination ability. In Substitute, Mike fails to discriminate something that is F (where F = grilled chicken) from something that is non-F (in this case – grilled soya chunks). We can generalize the example by presuming that Mike has never heard of soya chunks. Let`s further presume that Judith tells all of her and Mike`s mutual friends what happened at dinner, and they decide to pull an elaborate deception by deluding Mike into thinking that soya chunks are actually a premium kind of chicken. As a result, he starts to serve soya chunks in his restaurant and forms new false beliefs *about* roasted chicken. For example, that roasted chicken should be soaked before cooking, or that roasted chicken should be rehydrated before grilling. On the other hand, it would be strange to suggest that Mike has a deviant concept of roasted chicken, because he still possesses all his previous true beliefs about ‘roasted chicken’ and successfully discriminates things that are roasted chickens from all things that are not soya chunks³⁸.

³⁸ I think something similar happens in Burge`s brisket example – one`s social environment can shape one`s BRISKET concept in a way that allows only for partial understanding.

At this point, an internalist can question my argument on grounds of it being too strong. After all, it can be argued that Substitute relies on perceptual illusion or that it involves a convincing fake. From there, an internalist can generalise the rebuttal saying that Ability-CM would deny us possession of perceptual concepts because we can stipulate that all perceptual experiences can be convincingly faked³⁹. From there, the internalist can argue that a feasible option is to reinterpret Ability-CM as an even weaker condition, namely:

Ability CM*: S possess a concept C if and only if S possesses infallible knowledge how (or some other cognitive ability) to employ C correctly relativised to non-deceptive cases.

Such an interpretation of CM would be more plausible, if it was indeed available for the internalist. First, one has to take into account that other forms of internalism (e.g. internalism about knowledge or justification) have been having notoriously hard times dealing with the indistinguishability of good and bad cases⁴⁰. The reason for that is that such differences do not reside nowhere near the mental. Further, Ability CM* only confirms that by introducing the condition that infallibility should be regarded as relative to non-deceptive cases. However, excluding convincing fakes requires an external, environmental restriction on Ability CM that is not at the internalist's disposal.

The presumed phenomenal indistinguishability of the good and the bad case is just another reason why internalists should find a way around Substitute. If there is any internally accessible difference between Mike in the convincing fake scenario and his counterpart, Mike* who is not presented

³⁹ This line of defence was suggested to me by Johan Heemskerk. I would like to stress out that the plausibility of the argument will depend on the particular views about perceptual contents that the internalist is ready to endorse. However, I will not discuss this in detail here.

⁴⁰ Take BIV cases, on an internalist account Jane and JaneBIV have the exact same justification for believing that they have hands (for an excellent reconstruction of Cohen's 'new evil demon problem' see Srinivasan 2020, 406-407).

with a fake, but enjoys grilled chicken, the internalist has to account for that⁴¹.

A further challenge for Ability-CM was already raised by Burge's Arthritis case. It is not clear if, however we interpret it, a CM condition would allow for lack of scientific knowledge about the referent of the concept in question. Even if we find a way around Substitute, Mike may still reject "a trivial analyticity"⁴² about CHICKEN, like $p =$ "Chicken is *gallus domesticus*". Mike would probably also fail to understand sentences like s_2 "There is a *gallus domesticus* inside this dish". However, a zoologist can say that the linguistic expression 'gallus domesticus' still corresponds to the concept CHICKEN. Thus, the CM condition raises the question if Ability-CM does not also require too vast knowledge *about* a particular concept, its referent and its linguistic use.

To deny concept possession to Mike in either in these two scenarios would mean to deny it to too many subjects on too many occasions. Individualists are afraid that allowing incomplete understanding would imply that one does not understand her own thoughts. It turns out that the endorsement of ITC and the acceptance of the CM condition are actually what implies such a conclusion - Mike neither understands his thoughts (because according to CM he does not possess the concept ROASTED CHICKEN) nor he understands what he is saying. Therefore, holders of ITC should defend themselves against the same charge that they have put forward. It seems that we are not perfect thinkers, nor perfect speakers and, if ITC is indeed

⁴¹ Insisting that Ability CM is too strong because it introduces a discrimination ability will also not work, at least not for Wikforss' project. After all, she refuses to take the 'narrow content' way out of Burge's challenge in order to not sever the traditional link between thought-content and truth-conditions and to avoid the fragmentation of concept and reference (Wikforss 2001, 218). Other traditional internalist criteria for correctness like consistency with one's other beliefs will also not be applicable because, while they would secure that conceptual content is determined individualistically, they would still create a chasm between concepts and referents. As I agree with all of Wikforss' criticisms of traditional internalist responses to Burge, I will have nothing more to say about them.

⁴² Which I remind, according to Wikforss, would be exemplary of conceptual error (2001, 231).

correct, it seems that we actually fail to possess a lot of the concepts we operate with.

At this point, as pointed out to me by an anonymous reviewer, an internalist can object against Substitute by suggesting that Mike does not make a *competence* error, but a *performance error*. Such distinction would imply that a performance error is due to the external conditions in which the judgment is produced and that it does not require that Mike revises his ROASTED CHICKEN concept. If he was to make a competence error on the other hand, it would be related to his conceptual grasp of ROASTED CHICKEN and it would have suggested conceptual revision after Mike accepts that he made a mistake.

The answer to the internalist's objection consists of two parts. First, it is not clear at all that Mike is not expected to make a revision in regards to his ROASTED CHICKEN concept. Maybe he will adopt at least one new belief regarded to ROASTED CHICKEN, namely that roasted chickens are not the only thing that taste *like that*. If, before trying Judith's dish, he held the belief that a necessary and sufficient condition for something to be a ROASTED CHICKEN is to *taste like that*, then he probably would abandon this belief after trying the soya chunks (if we presume that Judith does inform him of the nature of his dish). Should not we suppose then that he has actually revised his ROASTED CHICKEN concept? Further, an all-out distinction between competence errors and performance errors may prove unavailing for the internalist. As I pointed out there can be cases of performance error where no competence error has been made. However, it is questionable if we can discuss *pure* competence errors without the presence of performance errors. After all, it is Bert's performance error in the Arthritis case that motivates Wikforss to suggest that he does not possess the concept ARTHRITIS. In Wikforss' own words: "Bert makes a conceptual error *when he utters* 'I have arthritis in my thigh'" (2004, 288).

The final problem encountered by Ability-CM concerns conceptual disagreement⁴³. If all subjects possess full understanding about the concepts they operate with and are, indeed, such perfect speakers, then how *can* we

⁴³ Wikforss is actually fully aware of that (2001, 227), but does not seem to provide a remedy to it apart from the appeal to individualists to not give up on reference and to not accept the conceptual and referential fragmentation (Ibid., 218, 226, 231)

account for instances where misunderstanding or disagreement arises? It looks like Wikforss' suggestion is that all conceptual differences appear, in the end, to be nothing more than differences in adopted theories (e.g. the actual and counterfactual theory of what 'arthritis' refers to in Burge's scenario) and not actual conceptual errors (Sawyer 2003, 270). However, if to be able to grasp a concept C, S has to satisfy the CM condition (Sawyer 2003, 273) the emerging figure of the perfect speaker will oblige the individualist to explain why no one ever makes a conceptual error⁴⁴. Furthermore, if every disagreement spurs from a difference of theories, and every theory has a chance to be proven correct in the future (Wikforss 2001, 225), ITC turns out to be enfolded in arbitrariness and relativity. It would be very difficult to point out clear cut criteria which should be adopted to distinguish true from false claims in an argument. Each attempt to outline such criteria would require that both sides in the argument talk about the same thing (i.e. grasp the same concept) and that according to said criteria one of the speakers is right and one of them is wrong.

Finally, even if I am right that Ability-CM suggests that concept possession requires that we are perfect thinkers, there may yet be a reader who remains unpersuaded by my arguments that such a condition has to imply that we are also perfect speakers. I would like to address this worry one last time and try to sway this reader to agree with me. While there may not be a one-to-one correspondence between conceptual and linguistic errors, as I already admitted above, a couple of things should be pointed out about conceptual errors⁴⁵: i) A conceptual error may not be *merely* a performance error, but it *is* a performance error (of a sort)⁴⁶; ii) evaluation and assessment of performance errors requires that the one who is being evaluated *performs*; iii) in the case of concept possession a performer is manifesting their ability *qua* the use of language. Thus, we may not be required to

⁴⁴ Either that, or in all cases of disagreement we should deny concept possession to both sides of the argument. However, this takes us back to my first objection because it seems that very few individuals would turn out to grasp any concepts at all.

⁴⁵ That is: about conceptual errors on an internalist account that endorses Ability-CM (see footnote 37 for a different problem with a different brand of internalism.)

⁴⁶ Remember that Ability CM requires correct application and infallible discrimination.

be perfect speakers in a loose sense (we are allowed to make purely linguistic errors), but we are required to be ones in a more strict and troubling sense (we are not allowed to apply the *word* ‘arthritis’ to ailments of the thigh). Avoiding the Davidsonian solution comes with a cost: bringing back reference suggests bringing back meaning, which in turn demands enforcing a (contingent) conceptual - linguistic relation.

All objections that were outlined above have the same consequence – even if incomplete understanding seems counterintuitive at first glance, to assume the opposite, expressed by the CM condition, comes with a heavy price to pay. ITC implies that we are perfect thinkers and perfect speakers and that, when we enter disagreement, we just talk about different things and follow different theories. On the other hand, the only implication of ETC and the rejection of P3 is that sometimes we are capable of ‘losing the keys to our cognitive home’ (so to say). Even if this is counterintuitive, no real arguments which are able to affirm that concept mastery is a necessary condition for grasping a concept were presented by the holders of ITC.

The rejection of P3 comes with an interesting consequence. Let’s reassess AIU and see what follows from the negation of the CM condition:

AIU (Redacted)

- P1: The unifying principle of reference can be used to justify the transition from externalism in regards of reference to ETC.
- P2: Arguments in favour of externalism in regards of reference often rely on cases where a subject S possesses only incomplete understanding about a concept C.
- C1: ETC would also require that we allow for concept possession in cases of incomplete understanding.
- P3: It is possible that there are cases where we do not completely understand our own thoughts.
- C2: The possibility of incomplete understanding can be used to ground reference externalism.
- C3: The unifying principle of reference can be used to ground ETC.

As per Wikforss’ own admission, the premise that incomplete understanding about our own concepts is possible grounds reference externalism. I showed that reference externalism is supported by something different than the

unifying principle, thus the principle can be true, and can “be used to defend the move from reference externalism to content externalism” (Wikforss 2004, 292). Therefore, per the redaction of Wikforss’ own argument, ETC is well-grounded and immune to individualists’ attacks.

Funding

My research was funded by the "Sofia University Marking Momentum for Innovation and Technological Transfer" (SUMMIT) project and I would like to express my gratitude for that.

Acknowledgements

I am very grateful to Rosen Lutskanov, Johan Heemskerk, Dimitar Elchinov and two anonymous reviewers for all their interesting and important comments on an earlier draft of this paper.

References

- Berker, Selim. 2008. “Luminosity Regained.” *Philosophers` Imprint* 8 (2): 1–22, <http://hdl.handle.net/2027/spo.3521354.0008.002> (accessed July, 22 2023).
- Burge, Tyler. 1979. “Individualism and the Mental.” *Midwest studies in philosophy* 4 (1): 73–121, <https://doi.org/10.1111/j.1475-4975.1979.tb00374.x>
- Burge, Tyler. 1986. “Intellectual Norms and Foundations of Mind.” *Journal of Philosophy* 83: 697–720, <https://doi.org/10.2307/2026694>
- Casalis, Séverine. 2004. “The concept of dyslexia”. In *Handbook of Children`s Literacy*, edited by T. Nunes & P. Bryant, 257–73. Springer Dordrecht, https://doi.org/10.1007/978-94-017-1731-1_15
- Davidson, Donald. 1987. “Knowing one’s Mind”. *Proceedings and Addresses of the American Philosophical Association* 60 (3), 441–58. <https://doi.org/10.2307/3131782>
- Rabin, Gabriel. 2020. “Toward a Theory of Concept Mastery: The Recognition View.” *Erkenntnis* 85 (3), 627–48. <https://doi.org/10.1007/s10670-018-0040-6>
- Rowland, Mark, Lau, Joe & Deutsch, Max. 2020. “Externalism about the mind.” In *The Stanford Encyclopedia of Philosophy* (Winter 2020 Edition), edited by Edward N. Zalta, Last updated December 10, 2020. <https://plato.stanford.edu/entries/content-externalism/>
- Sawyer, Sarah. 2003. “Conceptual Errors and Social Externalism.” *Philosophical Quarterly* 53 (211), 265–73, <https://doi.org/10.1111/1467-9213.00311>

- Sawyer, Sarah. 2015. "Internalism and Externalism in Mind." In *The Bloomsbury Companion to Philosophy of Mind*, edited by J. Garvey, 133–50. London: Bloomsbury Publishing. DOI: 10.5040/9781474244480.ch-007
- Sawyer, Sarah. 2018. "Subjective Externalism." *Theoria* 84 (1), 4–22, <https://doi.org/10.1111/theo.12137>
- Sawyer, Sarah. 2021. "Concepts, Conceptions and Self-Knowledge." *Erkenntnis*, 86, 237–54, <https://doi.org/10.1007/s10670-019-00109-2>
- Srinivasan, Amia. 2020. "Radical Externalism." *Philosophical Review*, 129 (3), 395–431, <https://doi.org/10.1215/00318108-8311261>
- Wikforss, Åsa. 2001. "Social Externalism and Conceptual Errors." *The Philosophical Quarterly* 51 (203), 217–31. <https://doi.org/10.1111/j.0031-8094.2001.00225.x>
- Wikforss, Åsa. 2004. "Externalism and Incomplete Understanding." *The Philosophical Quarterly* 54 (215), 287–94, <https://doi.org/10.1111/j.0031-8094.2004.00352.x>
- Williamson, Timothy. 1996. "Cognitive Homelessness." *The Journal of Philosophy* 93 (11), 554–73, <https://doi.org/10.2307/2941049>
- Williamson, Timothy. 2000. "*Knowledge and Its Limits*." New York: Oxford University Press, <https://doi.org/10.1093/019925656X.003.0012>
- Williamson, Timothy. 2016. "Abductive philosophy." *Philosophical Forum* 47 (3–4), 263–80, <https://doi.org/10.1111/phil.12122>

ISSUES ON THE (IM)POSSIBLE X

June 25-27, 2024 (Bratislava, SLOVAKIA)

KEYNOTE SPEAKERS

Barbara Vetter

(Freie Universität Berlin)

Dolf Rami

(Ruhr-Universität Bochum)

Manuel García-Carpintero

(Universitat de Barcelona)

We invite submissions for a 30-minute presentation, discussion excluded.

The conference focuses on analytic philosophy. The areas of interest include, but are not limited to modal metaphysics, metaphysics of fiction, hyperintensionality, metaphysics of responsibility and agency, modal epistemology, and metaontology.

An abstract of approximately 500 words should be prepared for blind review and include a cover page with the full name, institution, and contact information. Abstracts can be submitted in pdf or doc(x) and should be sent to impossible@savba.sk.

Informal queries: impossible@savba.sk

Webpage: www.metaphysics.sk/issuesontheimpossible/

Deadline for submission: **February 29, 2024**

Notification of acceptance: **March 31, 2024**