BELGRADE PHILOSOPHICAL ANNUAL/
FILOZOFSKI GODIŠNJAK 32/2019
Institute of Philosophy, Faculty of Philosophy,
University of Belgrade
Belgrade, Čika Ljubina 18–20

Belgrade

Year XXXII

YU ISSN 0353-3891

UDK-1

Belgrade Philosophical Annual online at http: //www.f.bg.ac.rs/bpa/index.html

Printed by
Službeni glasnik, Belgrade

This issue is financially supported by
Ministry of Education, Science and Technological Development of the Republic of Serbia

The statement on publication ethics can be found at the journal website http://www.f.bg.ac.rs/bpa

BELGRADE PHILOSOPHICAL ANNUAL 32/2019

MORAL PSYCHOLOGY

Guest editor: Voin Milevski

MORAL PSYCHOLOGY

Alexander Miller
University of Otago, New Zealand
Kirk Surgener
University of Warwick, UK

Original Scientific Paper
UDC 17.03:161/164"19/20"
17.022.1:161/164

ECUMENICAL EXPRESSIVISM AND THE FREGE-GEACH PROBLEM*

Abstract: *A background assumption of much of 20th century and recent metaethics and moral psychology is that moral judgements either express beliefs rather than desire-like attitudes or express desire-like attitudes rather than beliefs. In a recent series of papers and a monograph, Michael Ridge seeks to reject this assumption, and thereby to steer the focus of metaethical debate away from the Frege-Geach problem. In particular, Ridge claims that we can formulate "ecumenical" views on which moral judgements express both beliefs and desire-like attitudes, and that his own favoured metaethical position – Ecumenical Expressivism – can use the resources of cognitivism to provide a relatively straightforward solution to the Frege-Geach problem. In this paper we argue that Ridge's Ecumenical Expressivist response to the Frege-Geach problem is inadequate and explore the consequences of this inadequacy for our outlook on moral psychology.*

Keywords: *expressivism, cognitivism, Frege-Geach problem, ecumenicism*

1. Cognitivism and Expressivism

A traditional way of drawing the distinction between cognitivist and expressivist accounts of moral judgement characterizes cognitivists as holding that moral judgements express beliefs (and not desire-like attitudes) and expressivists as holding that moral judgements express desire-like attitudes (and not beliefs). Alternatively, the two positions could be summarised as follows:

> **Cognitivism:** For any moral sentence M, M is conventionally used to express a belief (and not a desire-like attitude).
>
> **Expressivism:** For any moral sentence M, M is conventionally used to express a desire-like attitude (and not a belief).

* For comments and discussion we're grateful to Finn Butler, Xin Cui, Ramon Das, Kent Hurtig, Richard Joyce, Simon Keller, Ed Mares, Alan Millar, Peter Milne, Glen Pettigrove, Nathan Sampson, Peter Sullivan, Justin Sytsma, Alan Weir, Camlo Woods, Crispin Wright, and seminar audiences at the University of Otago, the University of Stirling, and Victoria University of Wellington. For helpful comments on a distant ancestor of the current paper, thanks to Guy Fletcher and Neil Sinclair. Work on the paper progressed while Miller was visiting the Centre for the Study of Perceptual Experience at the University of Glasgow in June 2019: thanks to Fiona MacPherson for the invitation to visit the Centre.

A central problem for expressivism, thus characterised, is the Frege-Geach Problem, the problem of accounting for the meanings of moral sentences as they appear in unasserted contexts such as the antecedents of conditionals (Geach 1960, 1965).[1] Although the problem has been tackled by leading proponents of expressivism such as Allan Gibbard (1990, 2003) and Simon Blackburn (1984, 1993, 1998) it is fair to say that the solutions offered have not been convincing (see Schroeder 2008a and Miller 2013, chapters 4 and 5). In a series of articles (2006, 2007 2008, 2009) and recent monograph (2014), Michael Ridge has developed a novel form of expressivism, Ecumenical Expressivism, according to which moral judgements express *both* beliefs and desire-like attitudes, and argued that Ecumenical Expressivism enables a relatively straightforward solution to the Frege-Geach Problem.[2] Our main aim in this paper is to challenge Ridge's claim that Ecumenical Expressivism solves the Frege-Geach Problem. We proceed as follows. In §2 we give a very brief reminder of the Frege-Geach Problem. For illustrative purposes that we shall draw on later, we also recap the 1984 solution to the problem developed by Simon Blackburn and the main reason that Blackburn's solution fails. Following this, in §3 we explain Ridge's distinction between Ecumenical Cognitivism and Ecumenical Expressivism. In §4 we briefly outline how Ridge's Ecumenical Expressivism claims to solve the Frege-Geach Problem, before outlining, in the next four sections, a series of challenges to that solution. We set out our main conclusion and draw some broader morals in §9.[3]

2. Blackburn's Quasi-Realist Expressivism and the Frege-Geach Problem

The fundamental expressivist ideas are that we give an account of the meaning of a sentence in terms of the state of mind that it expresses and that

1 As Schroeder (2008a) rightly points out, the problem is much more general than simply dealing with the case of conditionals and concerns a family of issues surrounding compositionality in general: so the problem concerns the expressivist's capacity to preserve moral reasoning in general and not just e.g. moral modus ponens.

2 "Non-Ecumenical Expressivism" is thus the view that moral judgements express desire-like attitudes but not beliefs.

3 For the most part, for the purposes of evaluating Ridge's solution to the Frege-Geach Problem we focus on the simpler forms of Ecumenical Expressivism broached in his 2006 and 2008: the solution to the Frege-Geach Problem offered in Ridge 2014 is essentially the same as that offered in the earlier articles, with the additional complexities about normative perspectives, "admissible ultimate standards of practical reasoning" and "negative thinking" introduced in the 2014 Ecumenical Expressivist account playing (as far as we can see) no essential role in the attempt to defuse the Frege-Geach Problem. Likewise, we do not concern ourselves with the question as to whether expressivism is best framed as a thesis in semantics or (as Ridge now prefers) in metasemantics. As Ridge himself notes (2014: 137–38), the philosophical work that the expressivist has to carry out to deal with the Frege-Geach problem is effectively the same irrespective of whether it is couched as a view in first-order semantics or as a view in metasemantics.

in the case of a moral sentence such as "Murder is wrong" the relevant state of mind is a non-cognitive attitude of disapproval of murder: B!(murder).[4] These ideas, however, leave the expressivist with a problem. While it is plausible to think of the meaning of "Murder is wrong" as it appears in an asserted context such as e.g.

(1) Murder is wrong in terms of B!(murder), it is difficult to see how this account can be extended to cover the appearance of "murder is wrong" as it appears in an unasserted context such as the antecedent of (2):

(2) If murder is wrong then getting Peter to murder people is wrong, since someone sincerely asserting (2) needn't have an attitude of disapproval towards murder (or indeed towards getting Peter to murder people) – think of how those who approve of helping the aged can still sincerely utter "If helping the aged is wrong then getting Peter to help the aged is wrong". If this extension turns out not to be possible it looks like the inference from (1) and (2) to:

(3) Getting Peter to murder people is wrong will be vitiated by a fallacy of equivocation, since "Murder is wrong" will have different meanings as it appears in (1) and in the antecedent of (2). And this is highly problematic, as the inference is an instance of Modus Ponens, a valid inference form.[5] This is the Frege-Geach Problem, and the challenge to the expressivist is therefore to give an account of the contribution made by the meaning of a moral sentence to the meaning of a more complex sentence in which it appears in terms of the state of mind it expresses when used in an asserted context, in such a way that intuitively valid inferences involving it are not impugned (by, for instance, the commission of fallacies of equivocation).

It will be useful later to contrast Ridge's attempted solution with that attempted by Blackburn in his 1984. To cut to the chase, Blackburn proposes to understand the meaning of a conditional such as (2) above in terms of a higher-order attitude of approval towards moral sensibilities that combine

---

4 Ridge characterises expressivism as a form of "ideationalism", where "Ideationalism maintains that facts about the semantic contents of meaningful items in a natural language are constituted by facts about how those items are conventionally used to express states of mind" (2014: 107). For an account of the philosophical motivations for expressivism – in metaphysics, epistemology and moral psychology – see chapters 3–5 in Miller (2013).

5 Notice that it will not do for the expressivist to simply accept that this aspect of moral discourse is in bad faith: as we noted above the problem in this area extends to most of moral reasoning. Going down this road would leave the expressivist with an account of the meaning of positive, atomic, moral statements but not much else. At this point it is unclear why developing expressivism is preferable to simply adopting an error theory.

disapproval of murder with disapproval of getting Peter to murder people: schematically, H! [B! (Murder); B! (Getting Peter to murder people)]. If we now think of the overall state of mind of someone who accepts (1) and (2) but rejects (3) we can see that this will consist of disapproval of murder together with approval of combining disapproval of murder with disapproval of getting Peter to murder people, but will lack disapproval of getting Peter to murder people. Someone with this state of mind will be prey to a kind of incoherence: he "has a fractured sensibility which cannot itself be an object of approval" (1984: 195), and this allows us to capture the idea that the inference from (1) and (2) to (3) is valid.

This attempt at solving the Frege-Geach Problem was criticised shortly after its publication by Crispin Wright:

> Anything worth calling the validity of an inference has to reside in the inconsistency of accepting its premises but denying its conclusion. Blackburn does indeed speak of the 'clash of attitudes' involved in endorsing the premises of the modus ponens example, construed as he construes it, but in failing to endorse the conclusion. But nothing worth regarding as inconsistency seems to be involved. Those who do that merely fail to have every combination of attitudes of which they themselves approve. That is a *moral* failing, not a logical one (Wright 1988: 25).[6]

Blackburn's 1984 solution thus fails to capture the *logical* validity of the inference from (1) and (2) to (3). However, the key thing to note is that although it fails for the reason set out by Wright, it is nonetheless a genuine attempt to speak to the Frege-Geach worry about equivocation, since the contribution of "Murder is wrong" to the meaning of the conditional (2) is given in terms of the very same state of mind – B! (murder) – that gives its meaning in (1). This is a point we'll return to later.

3. Ecumenical Views

According to ecumenical views of moral judgement, moral judgements can be regarded as expressing *both* beliefs and desire-like attitudes: a moral sentence M is conventionally used to express *both* a belief and a desire-like attitude. This does not, however, lead to a collapse of the distinction between cognitivism and expressivism. According to Ridge, a version of this distinction survives the move towards ecumenicism:

> Ecumenical cognitivism allows that moral utterances express both beliefs and desires and insists that the utterances are true if and only

6 See also Hale (1986) and Hale (1993). For a useful extension of the sort of objection developed by Wright and Hale, see Van Roojen (1996).

> if one of the beliefs expressed is true. Ecumenical expressivism also allows that moral utterances express both beliefs and desires but denies that a moral utterance is *guaranteed* to be true just in case the belief(s) it expresses is (are) true (2006: 307–8, emphasis added).

And again:

> So long as the belief expressed by a moral utterance *is not semantically guaranteed* to provide the truth-conditions for the utterance, the fact that the belief expressed *contingently* provides the truth-conditions for the token utterance is consistent with expressivism as characterized here (2006: 311–312, emphases added).[7]

The distinction between cognitivism and expressivism within the ecumenical framework is thus recast as follows:

> **Ecumenical Cognitivism**: a moral judgement M expresses both a belief and a desire-like attitude, and, as a matter of semantic and conceptual necessity, M is true iff the belief expressed is true.
>
> **Ecumenical Expressivism**: a moral judgement M expresses both a belief and a desire-like attitude, but it is not semantically or conceptually necessary that M is true iff the belief expressed is true.

The Ecumenical Cognitivist assigns a certain logical priority to belief: which of an agent's judgements count as moral will be determined by the type of belief with which moral judgements necessarily co-vary; for example, a version of Ecumenical Cognitivism which took the beliefs in question to be beliefs about maximising utility would imply that the agent's moral judgements are those about the maximisation of utility. In contrast, although the Ecumenical Expressivist would regard moral judgements as expressing beliefs as well as desire-like attitudes, on this type of account logical priority would be assigned to the desire-like attitudes rather than the beliefs. For example, on the toy ("Plain Vanilla") version of Ecumenical Expressivism that Ridge sometimes uses in explaining the position:

> Normative utterances express (a) a speaker's approval [disapproval] of actions in general insofar as they have a certain property, and (b) a belief which makes anaphoric reference to that property (the one in virtue of which the speaker approves [disapproves] of actions in general) (2008: 55).

Consider a utilitarian speaker ("Jeremy"). Jeremy's judgement that X is right expresses (a) an attitude of approval towards actions insofar as they maximise utility and (b) a belief that X maximises utility. Which of Jeremy's judgements count as moral judgements will be determined by the characteristics towards which he takes the moral attitude of approval: since he takes this attitude

7 See also (2008: 54, 55, 59) for further use of "guarantee", "semantic guarantee" and so on.

towards actions which maximise utility, his moral judgements will be those judgements which express beliefs about utility maximisation.

Note that it is the Ecumenical Cognitivist's commitment to the semantic and conceptual necessity of the biconditional relationship between moral judgement and the type of belief assigned priority in the account which leaves it susceptible to Moorean "open question" style worries. Although the Ecumenical Expressivist may well posit a biconditional relationship between moral judgements and certain sorts of belief, that this relationship holds will be a matter of first-order normative theory:

> Given deflationism about truth and truth-aptness, the expressivist might hold that moral utterances are truth-apt but deny that their truth-conditions *necessarily* are provided by the beliefs they express. [T]he expressivist might argue that whether an agent's belief provides the truth-conditions for her utterance will be a substantive first-order question and not a question to be settled by metaethical theorizing (2006: 316, emphasis added).

Again

> [E]ven if normative utterances do express beliefs, as the Ecumenical Expressivist insists, they do not express beliefs which are such that the utterance is *semantically guaranteed* to be true just in case the belief is true (2008: 55, emphasis added).

Since the Ecumenical Expressivist does not view the relationship between the relevant type of belief and moral judgement to hold as a matter of semantic and conceptual necessity, he apparently escapes having to deal with "open question" style considerations.

And note, finally, that the Ecumenical Expressivist view leaves open the possibility of a kind of variability in what constitutes moral judgement. While Jeremy's moral judgements are keyed to utility in virtue of his attitude of approval towards utility maximising actions, Alvin's moral judgements may be keyed to a different characteristic in virtue of his attitude of approval being directed towards actions which instantiate it:

> Just what the relevant property is can vary from one speaker to the next. I might approve of actions insofar as they promote happiness, while you might approve of actions insofar as they are in accordance with God's will (2008: 55).

Thus, it may be that Alvin's judgement that X is right expresses (a) an attitude of approval towards actions insofar as they accord with God's will and (b) a belief that X accords with God's will.

4. Ridge's Solution

Ridge – conscious of the problem which undermined Blackburn's attempts at solving the Frege-Geach problem – articulates a constraint which any expressivist account has to meet:

> **Inconsistency Constraint:** the account must explain why someone who accepts the premises of a valid argument involving moral terms, but who denies the conclusion, is making a *logical* mistake. This inconsistency must be logical, rather than the pragmatic inconsistency exemplified by "Moore's paradox" style sentences, e.g. "I believe that P, but not-P" (see Ridge 2006: 313).

Since the expressivist has not – prior to solving the Frege-Geach Problem – earned the right to think of moral judgements as true or false, Ridge works with a notion of valid argument designed to avoid begging any questions by assuming that moral judgements can be regarded as having truth-values:

> **Validity:** An argument is valid just in case any [logically] possible believer who accepts all of the premises but at one and the same time denies the conclusion would thereby be guaranteed to have inconsistent beliefs (Ridge 2006: 326, "logically " inserted).

We can see how Ecumenical Expressivism proposes to solve the Frege-Geach problem by focussing on the "Plain Vanilla" version outlined above, using our utilitarian speaker Jeremy as a representative believer. Suppose that Jeremy accepts premises (1) and (2) but rejects the conclusion (3). In virtue of accepting premise (1), Jeremy expresses the belief that murder maximises disutility; in virtue of accepting premise (2) he expresses the belief that if murder maximises disutility then getting Peter to murder people maximises disutility; in virtue of rejecting (3) he expresses the belief that getting Peter to murder people does not maximise disutility. He thus has straightforwardly inconsistent beliefs. So the argument is valid.

Nothing turns on Jeremy in particular. Suppose that Alvin accepts premises (1) and (2) but rejects the conclusion (3). In virtue of accepting premise (1), Alvin expresses the belief that murder clashes with God's will; in virtue of accepting premise (2) he expresses the belief that if murder clashes with God's will then getting Peter to murder people clashes with God's will; in virtue of rejecting (3) he expresses the belief that getting Peter to murder people does not clash with God's will. He thus has straightforwardly inconsistent beliefs. So, again, the argument is valid.

Ecumenical Expressivism thus exploits the fact that moral judgements express beliefs as well as desire-like attitudes to avoid the Frege-Geach Problem. In the remainder of the paper, we'll outline three problems that suggest that Ecumenical Expressivism fails to provide a convincing solution to the Frege-Geach Problem.

5. First Problem: Security Against Equivocation?

What guarantees that "Murder is wrong", as it appears in the antecedent of (2), has the same meaning as it has in the initial premise (1)? Recall that the truth-conditions of the beliefs about (dis)utility expressed by Jeremy's moral judgements are not semantically or conceptually guaranteed to be the truth-conditions of those judgements: this is what makes the view a form of Ecumenical Expressivism as opposed to Ecumenical Cognitivism. So the fact that beliefs about (dis)utility are expressed by Jeremy's acceptance of (1) and acceptance of (2) cannot on its own secure the univocity of "Murder is wrong" as it appears in those premises. In order to secure the argument against equivocation, note has to be taken in addition of the role played by the desire-like attitude expressed. Ridge's idea (see e.g. 2014: 152) is that this remains constant in the states of mind expressed by the acceptance of the premises and the rejection of the conclusion and that it is the *combination* of this attitude and the relevant beliefs about e.g. (dis)utility that guarantees univocity.[8] As a first pass, we can say that the hybrid states of mind Jeremy expresses in virtue of accepting (1) and (2) and rejecting (3) are:

(i) (Belief that murder maximises disutility, B!(actions which increase disutility))
(ii) (Belief that if murder maximises disutility then getting Peter to murder people maximises disutility, B!(actions which increase disutility))
(iii) (Belief that getting Peter to murder people does not maximise disutility, B! (actions which increase disutility))

We will now argue that this fails to secure the inference against equivocation. In order to secure univocity, the contribution of the antecedent ("Murder is wrong") to the meaning of the entire conditional (2) must be given by the state of mind expressed by the antecedent as it appears in the asserted context (1). In order to see how Ridge's account fails to do this, note first that in order for the belief that murder maximises disutility and the general sentiment *B!(actions which increase disutility)* to conjointly constitute a moral judgement they have to be related in some way: Ridge says explicitly (2008: 71) that normative judgement is constituted by there being a *link* between the relevant belief and desire-like attitude, and he also (2014: 195) refers to it as a "relational state".[9] (At a minimum, presumably, the belief and desire-like attitude need to be able to interact with each other in the psychological economy of the relevant agent). Suppose that the relevant relation is R. Then, the state of mind expressed in virtue of Jeremy's acceptance of (1) is

(i*) R(belief that murder maximises disutility, *B! (actions which increase disutility)*)

8 See also Schroeder (2009b: 197–8).

9 See also Schroeder (2013: 307–8).

In other words, the complex state of mind that consists in the belief that murder maximises disutility standing in the relation R to the general sentiment *B! (actions which increase disutility)*.

Likewise, the state of mind expressed in virtue of Jeremy's acceptance of the conditional (2) is

(ii*) R(belief that if murder maximises disutility then getting Peter to murder people maximises disutility, *B!(actions which increase disutility)*)

i.e. the state of mind that consists in the belief that if murder maximises disutility then getting Peter to murder people maximises disutility standing in relation R to the general sentiment *B! (actions which increase disutility)*.

Our key claim here is that since the state of mind contributed by "murder is wrong" to (ii*) is not (i*), Ridge fails to deal convincingly with the problem about equivocation. It is perhaps easiest to see this by reflecting on the fact that the state of mind (i*) is not a component of the state of mind (ii*) is the way in which, on Blackburn's 1984 account, the state of mind expressed by (1) is a component of the state of mind expressed by (2). Recall from §2 above that for Blackburn the state of mind expressed by (1) is

(i**) B! (murder)

while the state of mind expressed by (2) is

(ii**) H! [*B! (murder)*; B! (getting Peter to murder people)]

Here, the contribution of "murder is wrong" to the state of mind expressed by the conditional (*italicised*) is given by the very same state of mind expressed in the simple asserted context. This is not the case in Ridge's Ecumenical Expressivist account: the complex state of mind (i*) is not what "murder is wrong" contributes to (ii*). Hence Ecumenical Expressivism fails to secure univocity, and the security against equivocation required for a viable solution to the Frege-Geach problem is not provided.[10], [11]

---

10 It appears that the most Ridge can say is that "Murder is wrong" contributes the relation R, the belief that murder maximises disutility and the attitude B!(actions which increase disutility). On its own, this isn't sufficient to guarantee univocity: it is consistent with e.g. "Murder is wrong" contributing the state of mind R(B!(disutility causing actions), belief that murder causes disutility), and since we don't know whether R is symmetric, this may well not be the same state of mind expressed in virtue of Jeremy's acceptance of (1). The most that Ridge can legitimately say here is that "Murder is wrong" contributes R, the belief that murder maximises disutility, and B! (actions which increase disutility), but – crucially – not in a way that displays them as determinants of the state of mind expressed by (i*). (The argument of this section was sparked by a suggestive comment by Neil Sinclair, and deploys a strategy similar to that used in Sinclair (2011) against the account of sentential negation developed in Schroeder (2008b)).

11 On Ridge's account, how does acceptance of the premises in a moral modus ponens argument commit me to acceptance of the conclusion? An answer might be that in

6. Second Problem: Agnostics about First Order Nonconditional Matters

In order to outline this problem we'll work with the "Ideal Observer" version of Ecumenical Expressivism favoured in Ridge (2006). On this, an agent's judgement that e.g. X is morally required expresses (a) an attitude of approval towards actions insofar as they garner approval from a certain sort of ideal observer and (b) a belief that X would garner approval from that kind of ideal observer.

Ridge allows (2006: 334–336) that there are at least two ways in which a conditional statement can be accepted. Consider

> (B) If passive euthanasia is sometimes morally required then active euthanasia is sometimes morally required.

---

accepting the premises I express beliefs whose acceptance commits me to the belief expressed by the conclusion. But how do these beliefs commit me to the desire-like attitude expressed in accepting the conclusion? John Eriksson notes Mark Schroeder's suggestion (2009b: 198) that the key to this is the idea – noted above – that acceptance of any moral sentence containing e.g. "wrong" will for me express the same desire-like attitude. Eriksson argues against this that while this explains why someone who accepts the premises *has* the desire-like attitude prescribed by the conclusion, it fails to explain why someone who accepts the premises is *committed* to accepting the conclusion. He writes:

> [I]t seems more reasonable to think that the kind of attitude prescribed by the conclusion is a new attitude and not an attitude one has merely in virtue of accepting the premises. For instance, it seems conceivable that an agent accepts the premises yet fails to accept the conclusion, but if someone who accepts the premises already has the desire-like attitude prescribed by the conclusion, this seems impossible (2009: 15–16).

The obvious reply to this is that someone who has the desire-like attitude expressed by the conclusion need not have the belief it expresses, so that they needn't have the belief-desire pair possession of which would constitute acceptance of the conclusion. Eriksson objects that this misses the point, since:

> [First], it should be possible to accept the premises without thereby having the attitude expressed by the conclusion. Second, the objection turns on the fact that one does not necessarily have the belief expressed in the conclusion. However, it seems possible to have the belief but, for some reason or other, fail to acquire the desire-like state of mind expressed by the conclusion. This still seems to be something that Ridge's view rules out (2009: 16, n.26).

This strikes us as weak. Without additional argument, the unsupported assertion that it should be possible to accept the premises without thereby having the *desire-like attitude* expressed in the conclusion simply begs the question against Ridge. And the possibility that Eriksson mentions in his second point is not ruled out: someone who doesn't accept the premises may on Ridge's account be able to have the belief component of the conclusion without having the desire-like attitude. Eriksson's objection to Ridge thus seems to us to fail. Whether the variant "ecumenical" position he goes on to develop as an alternative to Ridge's is itself plausible is a matter for future discussion. Likewise for the "ecumenical" position developed in Toppinen (2013). (Note that Eriksson (2009) refers to an unpublished paper by Schroeder called "Finagling Frege": the point discussed appears to have appeared in print since in Schroeder (2009b), to which we refer above).

The standard way of accepting (B) involves having a state of mind that consists of an attitude of approval towards actions insofar as they garner approval from a certain sort of ideal observer together with a belief that if passive euthanasia (PE) sometimes garners the approval of that sort of ideal observer then so does active euthanasia (AE). Ridge admits that (B) may also be accepted by an agent who has suspended judgement about all first-order moral matters (i.e. someone who neither approves nor disapproves of actions):

> Here, I suggest that it is most plausible within the framework of Ecumenical Expressivism to understand such an agent as taking a stand against the approval of certain sorts of observers—those observers who would simultaneously approve of passive euthanasia but at one and the same time not also approve of active euthanasia, say. In the Ecumenical framework, this will amount to the agent's adopting a perfectly general noncognitive attitude, here an attitude of refusal—refusal to approve of an observer unless it has certain features and the belief that such features (once again we have a belief with anaphoric reference back to the content of a noncognitive attitude) preclude simultaneously approving of passive euthanasia while not also approving of active euthanasia (2006: 335).

Ridge notes a potential worry opened up by this sort of multiple realizability:

> The only problem, so far as the technical details of the solution to the Frege-Geach puzzle go, would arise if it were possible for someone to accept a conditional premise in the way characteristic of someone who is agnostic on all substantive nonconditional first-order normative claims, while at one and the same time accepting a nonconditional substantive first-order premise in the more standard way. For in this sort of case, if it were possible, the belief expressed in the major premise would not "hook up" logically in the right way with the belief expressed by the conditional premise to explain the validity of the argument (2006: 335).

Call this putative "bifurcated" moral agent "Sick Boy". Suppose that he accepts (A) and (B) but rejects (C):

(A) Passive euthanasia is sometimes required.
(B) If passive euthanasia is sometimes required then active euthanasia is sometimes required.
(C) Active euthanasia is sometimes required.

If such a "bifurcated" Sick Boy were possible this would frustrate Ridge's solution to the Frege-Geach problem: bifurcated Sick Boy would accept

(A) and (B) and reject (C) but would not thereby be guaranteed to have inconsistent beliefs, so that we would have a plainly valid argument that turned out not to be valid on Ridge's conception of validity. However, Ridge argues that bifurcated Sick Boy isn't in fact possible:

> [S]uch cases are not possible on the theory on offer here, properly understood. For if someone does have a normative outlook at all, as they must to accept an atomic judgment like passive euthanasia is right, then they can only count as making the relevant conditional judgment if they have the right sort of belief about that observer. Refusing to approve of certain sorts of observers can play a role in conditional (and other nonatomic) moral judgments only when someone lacks a normative outlook. *Once someone adopts a general normative stance by approving of a certain sort of observer, it is plausible to hold that this is dominant in determining their normative judgments, including their conditional judgments, and that they therefore simply do not count as judging, for example, that if passive euthanasia is right then so is active euthanasia unless they believe that the observer they take to be ideal would approve of the former only if he also approved of the latter* (2006: 335–336, emphasis added).

How plausible is Ridge's claim that there cannot be an agent who accepts nonconditional moral statements in the standard way and conditional moral statements in the manner of an agnostic about first order moral matters? It might well be true as a matter of empirical fact (or possibly even as a matter of psychological necessity) that the normative stance of the non-agnostic about first order nonconditional statements would be dominant and come into play in the agent's acceptance of conditional moral statements, but the crucial question is *whether this is so as a matter of logical necessity*: so long as bifurcated Sick Boy is logically possible, we have on Ridge's account a logically possible agent who accepts the premises of a moral modus ponens argument while rejecting the conclusion but who is not thereby guaranteed to have inconsistent beliefs.

Is bifurcated Sick Boy logically impossible? Let's think about his overall state of mind. In virtue of accepting (A), Sick Boy approves of actions insofar as they garner approval from a particular kind of ideal observer (call him I), and he believes that passive euthanasia sometimes garners approval from I. In virtue of rejecting (C), he approves of actions insofar as they garner approval from I but believes that active euthanasia does not sometimes garner approval from I. Putting these together we can say that Sick Boy approves of an observer (I) who sometimes approves of passive euthanasia without sometimes approving of active euthanasia. However, in virtue of his acceptance of the conditional (2) in the manner of an agnostic about first-order moral matters, he refuses to approve of an observer unless that observer has some feature which precludes sometimes approving of passive

euthanasia without sometimes approving of active euthanasia. *The most we can say about Sick Boy is that in approving of I he does something that he has a stance of refusing to do.* Plainly, this is a moral failing not unlike that of the agent who fails to have every combination of attitudes of which he himself approves. He fails to live up to his commitments. Agents who fail to live up to their commitments in this way are logically possible! Moreover, such agents commit no logical error: if there were some logical incoherence in failing to live up to one's commitments (in doing what you have a stance of refusing to do) Blackburn's solution to the Frege-Geach Problem would not have succumbed to the objection from Wright outlined in section 2 above. Since there is no logical incoherence in the idea of bifurcated Sick Boy, Ridge fails to dispatch the worry opened up by his concession that there are multiple ways in which a conditional moral statement can be accepted.

7. Third Problem: Variability Within a Single Agent

Recall from §3 above that on Ecumenical Expressivism it is possible for different speakers to make identical moral judgments in different ways. Reverting back to "Plain Vanilla" Ecumenical Expressivism, it may be that Jeremy's judgement that x is right expresses a complex state of mind consisting of a generalised attitude of approval *H!(actions which maximize happiness)* together with the belief that x maximizes happiness, while Onora's judgement that x is right expresses a complex state of mind consisting of a generalized attitude of approval *H!(actions which comply with the Categorical Imperative)* together with the belief that x complies with the Categorical Imperative.[12]

We might ask: if we can have this sort of variability between different speakers, why not within a single speaker at a single time with respect to different types of claim? For example, say that Dee is a utilitarian vis a vis some non-conditional claims but a Kantian vis a vis some conditional claims. Then suppose that Dee accepts (a) and (b) but rejects (c) in:

(a) x is right
(b) If x is right then y is right
(c) y is right.

Then Dee will have the following hybrid states of mind:

(a*) H!(things which maximize happiness); belief that x maximizes happiness.
(b*) H!(actions which comply with the Categorical Imperative); belief that if x complies with the Categorical Imperative then y complies with the Categorical Imperative.

12 For ease of exposition we here suppress mention of the relation which binds the belief and the attitude together in the complex state of mind: nothing turns on this here.

(c*) H!(things which maximize happiness); belief that y does not maximize happiness.

There is no inconsistency in Dee's beliefs, supplying a counterexample to Ridge's account of validity.

Ridge must therefore argue that an agent like Dee is logically impossible. Let's call the attitudes H!(things which maximize happiness) and H!(actions which comply with the Categorical Imperative) *normative perspectives*. Our question is therefore whether there is some logical or conceptual incoherence in the idea of someone occupying variable normative perspectives in the manner of Dee. What does Ridge have to say about this?

In his 2014 book Ridge introduces the notion of a normative perspective, where this is defined as the complete set of an agent's "emotionally tinged self-governing policies" (2014: 152) rather than in terms of a single generalised attitude of approval or disapproval. That an extension of this sort is required is shown by examples such as the conditional:

(E) If x is right then y is wrong.

The complex state of mind expressed when Jeremy accepts this will need to contain both a generalised attitude of approval and a generalised attitude of disapproval together with beliefs keyed to the characteristics which the attitudes are directed at:

(E*) belief that if x maximizes happiness then y maximizes unhappiness; {H!(actions which maximize happiness, B!(actions which maximizes unhappiness)}

The normative perspectives that Ridge speaks of in his 2014 are simply generalized versions of the set which forms the second component of (E*).

To return to our question: is an agent like Dee, occupying different normative perspectives vis a vis conditional and nonconditional statements, logically possible, so that equivocation in the beliefs relevant to the validity of an argument results in some valid arguments being deemed invalid? Considering a worry along these lines, Ridge writes:

> Given that an agent can at any given point in time have only one normative perspective this ensures that [there is no equivocation among] the beliefs relevant to testing the validity of the relevant arguments (2014: 152).

In Dee's case, the description of the single normative perspective that he occupies would presumably consist of the attitude *H!(things which maximize happiness)* and the attitude *H!(actions which comply with the Categorical Imperative)* together with some indication to the effect that the former kicks in when Dee is considering nonconditional statements while the latter kicks in when he is considering conditional statements. Presumably, equivocation

is avoided because the contents of the beliefs involved become disjunctive. In the example above Dee's beliefs will include: the belief that x either maximizes happiness or complies with the Categorical Imperative, the belief that if x maximizes happiness or complies with the Categorical Imperative then y maximizes happiness or complies with the Categorical Imperative, and the belief that y neither maximizes happiness nor complies with the Categorical Imperative. These beliefs are inconsistent as a simple matter of logic, so that the alleged counterexample of Ridge's account of validity is avoided.

However it turns out that this "solution" is only made possible because of a *stipulative definition* Ridge makes concerning "normative perspective":

> Another important feature of the view is that, by definition, a speaker will count as occupying at most one normative perspective at any given point in time. Whenever it seems that a speaker occupies more than one, the right thing to say is that his normative perspective is really the conjunction of what one might otherwise take to be his normative perspectives. This is simply how I am defining *normative perspective* here, as a term of art – they are by definition maximally general in this way (2014: 121).
>
> An agent can at any given point in time have only one normative perspective because normative perspectives are just defined as the totality of the relevant sorts of emotionally tinged self governing policies (2014: 152).

It follows from this that the "solution" to the Frege-Geach offered by Ridge is merely a trivial consequence of a stipulative definition: Ridge has simply defined "normative perspective" in such a way that normative perspectives are guaranteed to have a characteristic (non-variability in a single agent at a single time), a consequence of which is that in accepting the premises but rejecting the conclusion of a moral modus ponens argument the relevant agent has inconsistent beliefs. What Ridge owes us is some *non-ad hoc, substantive* reason for thinking that no logically possible believer can occupy variable normative perspectives in this way. Given that this has not been provided, we have not been given a compelling solution to the Frege-Geach Problem.

8. Schroeder's Objection

It might be worthwhile at this point to pause briefly in order to explain how our objection to Ridge's attempted solution of the Frege-Geach problem differs from an objection that has been developed by Mark Schroeder (Schroeder 2009a).

Schroeder's objection starts out from the observation that Ridge's 2006 account of moral sentences sees them as involving a kind of sentential anaphora. "Murder is wrong", for example, is held by Ridge to express (A)

a desire-like sentiment of disapproval towards action-types insofar as they possess a certain property and (B) a belief that murder possesses *that* property. The pronoun in (B) is anaphoric on the reference to the property in (A). Now consider the following:

> Superman flies.
> If Clark Kent flies then I'm a walrus. So,
> I'm a walrus.

This is truth-preserving but not logically valid: someone who isn't party to the substantive information that Superman and Clark Kent are the same man could rationally accept (a) and (b) and deny (c). Likewise for

> Superman – he flies.
> But Clark Kent – if he flies then I'm a walrus. So,
> I'm a walrus.

This is truth-preserving given the preferred interpretation of "Superman" and "Clark Kent", but for logical validity we require truth-preservingness in *any* model, not just in the preferred interpretation.

According to Schroeder the moral modus ponens argument is akin to these because seeing that the moral MPP argument is truth-preserving on Ridge's interpretation requires knowledge of the substantive assumption that moral sentences all express the same desire-like attitude. Without that assumption there is no guarantee that the belief expressed in the first premise of the moral MPP is the same as that expressed in the antecedent of the conditional second premise. So Ridge has not captured the logical validity of moral MPP and so has failed to solve the Frege-Geach problem "on the cheap".

Schroeder's objection is subtle and deserves more careful attention than we can give it here. However, it does seem to us that Schroeder's objection is somewhat narrower than that presented in some of the influential presentations of the Frege-Geach problem in its application to Blackburn's quasi-realism, such as Hale (1986, 1993) and Wright (1988). There the objection seems to be that Blackburn cannot frame the moral MPP argument in a way that satisfies some expressivist surrogate of the notion of truth-preservingess. The moral MPP argument on Blackburn's account doesn't do this because it is no better than an argument that equivocates and which has true premises and a false conclusion – and which is therefore *a fortiori* not truth-preserving (or possessed of a surrogate thereof). We see the objection we raised against Ridge above as concerning this more general worry: the moral MPP argument on Ridge's interpretation is not *even* truth-preserving (because of its failure to deal with the worries about equivocation) and is therefore not logically valid (since being truth-preserving is a necessary – though not sufficient – condition for logical validity). Whereas Schroeder's

worry is that on Ridge's account moral modus ponens arguments are truth-preserving but not truth-preserving in virtue of their form, our worry is that they are not truth-preserving at all. [13]

9. Conclusion

Overall, we can conclude that Ecumenical Expressivism does not offer a solution to the Frege-Geach problem that succeeds where the solutions offered by Non-Ecumenical Expressivism fail.

What lessons can we draw from this discussion for moral psychology in general? Ridge is committed to the Humean view that beliefs and desires are "distinct existences" (2014: 49–50). Abstracting a little from the specifics of our argument, what seems to be driving the problem for Ridge is this: to get the right kind of guarantee needed for a successful solution to the Frege-Geach problem you need a much tighter connection between the belief and desire-like elements posited than Ridge's account allows.[14] To put this into the context of the history of moral psychology, we can see now why one might be driven to posit a "besire"-friendly view, where moral judgements are taken to express unitary mental states with both desire-like and belief-like features.[15] Whatever the deficiencies of such a position at least the view earns a robust connection between desire-like and belief-like features through commitment to a non-Humean metaphysics of mental states. What we are suggesting is that Ridge cannot have his cake and eat it: without a more radical departure in our theory of motivation than he countenances a viable solution to the Frege-Geach problem will elude him. Alternatively, one could retain a commitment to a Humean theory of motivation but then the view will have no substantial advantage over other, non-ecumenical, versions of expressivism that allow for ethical statements to communicate descriptive information.[16] Thus the terrain of moral psychology is much more tightly constrained than in Ridge's vision.

---

13 This is not to say that Schroeder's objection to Ridge's account of formal validity is not a good one, just that it is not the most fundamental problem in the vicinity. In fact, Ridge attempts in his 2014 to extend his 2006 account of validity in a way that speaks to Schroeder's objection: see (Ridge 2014: 153–159). We remain neutral here on whether the developments introduced by Ridge succesfully deal with Schroeder's objection.

14 Although this has not formed part of our case here, we suspect similar considerations apply to Ecumenical Cognitivism as construed by Ridge, and its attempt to secure motivational internalism – again, the framework Ridge provides doesn't allow for a tight enough connection between the cognitive and the conative to do justice to the phenomenon in question.

15 See for instance Altham (1984). For discussion of the deficiencies of this kind of view, see Smith (1994).

16 For example, if you know that I morally approve of all and only actions that maximize the number of green things in existence, you will be able to infer from my calling an action right that I believe it will maximize the green things in existence. For a brief overview of views in this ballpark see van Roojen (2018).

This final consideration allows us to note that Ecumenical Expressivism's inability to succeed where Non-Ecumenical Expressivism fails should perhaps have been obvious from the start. In *Spreading The Word*, Blackburn wrote:

> We can see that it does not matter at all if an utterance is descriptive as well as expressive, provided that its distinctive meaning is expressive. It is the *extra import* making the term evaluative as well as descriptive, which must be given an expressive role. It is only if that involves an extra truth-condition that expressivism about values is impugned (Blackburn 1984: 169–70).

In effect, Blackburn is here countenancing the type of Ecumenical Expressivist view favoured by Ridge. It seems, then, that either Ridge has a simple solution to the Frege-Geach Problem that Blackburn somehow missed despite countenancing the possibility of the view or what Ridge takes be to be a simple solution to the Frege-Geach Problem is in fact no solution at all. The problems outlined above suggest that the latter is the case.

References

Altham, J. 1984: "The Legacy of Emotivism", in G. Macdonald and C. Wright (eds.) *Fact, Science and Morality* (Oxford: Basil Blackwell).

Blackburn, S. 1984: *Spreading the Word* (Oxford: Oxford University Press).

Blackburn, S. 1993: *Essays in Quasi-Realism* (Oxford: Oxford University Press).

Blackburn, S. 1998: *Ruling Passions* (Oxford: Oxford University Press).

Eriksson, J. 2009: "Homage to Hare: Ecumenicism and the Frege-Geach Problem", *Ethics* 120, 8–35.

Geach, P.T. 1960: "Ascriptivism", *Philosophical Review* 69, 221–225.

Geach, P.T. 1965: "Assertion", *Philosophical Review* 74, 449–465.

Gibbard, A. 1990: *Wise Choices, Apt Feelings* (Cambridge MA: Harvard University Press).

Gibbard, A. 2003: *Thinking How to Live* (Cambridge MA: Harvard University Press).

Hale, B. 1986: "The Compleat Projectivist", *Philosophical Quarterly* 34, 65–84.

Hale, B. 1993: "Can There Be a Logic of Attitudes?", in J. Haldane and C. Wright (eds.) *Reality, Representation and Projection* (Oxford: Oxford University Press).

Miller, A. 2013: *Contemporary Metaethics: An Introduction* (Cambridge: Polity Press)

Ridge, M. 2006: "Ecumenical Expressivism: Finessing Frege", *Ethics* 116, 302–36.

Ridge, M. 2007: "Epistemology for Ecumenical Expressivists", *Proceedings of the Aristotelian Society Supplementary Volume* LXXXI, 83–108.

Ridge, M. 2008: "Ecumenical Expressivism: The Best of Both Worlds?", in R. Shafer-Landau (ed.) *Oxford Studies in Metaethics* vol. 2 (Oxford: Oxford University Press), 51–76.

Ridge, M. 2009: "The Truth in Ecumenical Expressivism", in D. Sobel and S. Wall (eds.) *Reasons for Action* (Cambridge: Cambridge University Press).

Ridge, M. 2014: *Impassioned Belief* (Oxford: Oxford University Press).

Schroeder, M. 2008a: "What is the Frege-Geach Problem?", *Philosophy Compass* 3/4, 703–720.

Schroeder, M. 2008b: *Being For: Evaluating the Semantic Program of Expressivism* (Oxford: Oxford University Press).

Schroeder, M. 2009a: "Hybrid Expressivism: Virtues and Vices", *Ethics* 119, 257–309.

Schroeder, M. 2009b: *Noncognitivism in Ethics* (New York: Routledge).

Schroeder, M. 2013: "Tempered Expressivism", *Oxford Studies in Metaethics* 8, 283–313.

Sinclair, N. 2011: "Moral Expressivism and Sentential Negation", *Philosophical Studies* 152, 385–411.

Smith, M. 1994: *The Moral Problem* (Oxford: Blackwell).

Toppinen, T. 2013: "Believing in Expressivism", *Oxford Studies in Metaethics* 8, 252–282.

Van Roojen, M 1996: "Expressivism and Irrationality", *Philosophical Review* 105, 311–335.

Van Roojen, M. 2018: "Moral Cognitivism vs Non-Cognitivism" *Stanford Encylopedia of Philosophy.*

Wright, C. 1988: "Realism, Anti-Realism, Irrealism, Quasi-Realism", reprinted in *Saving The Differences* (Camb, MA: Harvard University Press 2003).

Marija Kušić
Departmant of Psychology
Laboratory for Research of Individual Differences
Faculty of Philosophy, University of Belgrade
Petar Nurkić
Departmant of Philosophy
Institute of Philosophy
Faculty of Philosophy, University of Belgrade

Original Scientific Paper
UDC 004.8: 17.018.21
004.85:[159.9:17

ARTIFICIAL MORALITY: MAKING OF THE ARTIFICIAL MORAL AGENTS

Abstract: *Artificial Morality is a new, emerging interdisciplinary field that centres around the idea of creating artificial moral agents, or AMAs, by implementing moral competence in artificial systems. AMAs are ought to be autonomous agents capable of socially correct judgements and ethically functional behaviour. This request for moral machines comes from the changes in everyday practice, where artificial systems are being frequently used in a variety of situations from home help and elderly care purposes to banking and court algorithms. It is therefore important to create reliable and responsible machines based on the same ethical principles that society demands from people. New challenges in creating such agents appear. There are philosophical questions about a machine's potential to be an agent, or moral agent, in the first place. Then comes the problem of social acceptance of such machines, regardless of their theoretic agency status. As a result of efforts to resolve this problem, there are insinuations of needed additional psychological (emotional and cognitive) competence in cold moral machines. What makes this endeavour of developing AMAs even harder is the complexity of the technical, engineering aspect of their creation. Implementation approaches such as top-down, bottom-up and hybrid approach aim to find the best way of developing fully moral agents, but they encounter their own problems throughout this effort.*

Keywords: *Artificial morality, artificial moral agents, machine learning, moral psychology, hybrid model*

1. Introduction

Artificial Morality is a new interdisciplinary field of research within Moral psychology and Machine engineering (i.e. Robotics). In the last decade, due to technological advances, it has been developing at an exponential rate.[1]

1 The work on this paper has been supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia through the project *Dynamic Systems in Nature and Society: Philosophical and Empirical Aspects* (No. 179041).

 DOI: 10.5937/BPA1932027K

Synonymously called Machine Ethics, Artificial Morality aims to create self-governing, ethical machines that can "function in an ethically responsible manner", that is, machines capable of making autonomous decisions that are in accordance with the society's norms and moral standards (Anderson & Anderson, 2007, pp 15; Allen, Smith & Wallach, 2005, pp 149). To enable morally functioning machines, Artificial Morality considers different ethical principles or learning procedures that govern human behaviour and enable them to act as moral agents. These governing principles are then algorithmically formalized and implemented in machines, thus creating new artificial moral agents. (Anderson & Anderson, 2007, pp 15; Misselhorn, 2018, pp 161).

Artificial Morality can be classified into the subfields of both computer science (more closely, artificial intelligence) and moral psychology (or moral philosophy), predominately because of its eclectic, interdisciplinary approach (Yampolskiy, 2013, pp 389). As a starting point in creating morally competent agents, it uses the achievements of cognitive science and ethics. The main task, when establishing the basic functioning principles of machines, is the abstraction of elements of human moral reasoning and behaviour (Malle, 2015, pp 243) or formalization of ethical principles into computer algorithms (Yampolskiy, 2013, pp 389).

This paper will try to exhibit the complex structure of the Artificial Morality field by dividing it into three main parts (or problems).[2] The first one is the conceptual problem of machines as moral agents, more closely, the mere possibility of machines being moral agents equivalent to humans. This problem is a philosophical one. It grips the normative nature of the field – modality of moral machines – best conceptualized in the question "can machines be moral agents?". Answering this question requires considering the components of moral agency and realizable ways in which machine behaviour can come close to human behaviour.

The second part considers the descriptive, psychological problem that comes after resolving the previous one, namely, the problem of social acceptance of autonomous machines. Moreover, apart from the machines' ability to "function in an ethically responsible manner", it is important to know whether they are going to be accepted and trusted as such autonomous agents, and what will make them more trustworthy in the eyes of society. The problem of interest is how to make technically functional moral machines to also be socially functional agents. In other words, Artificial Morality also deals with the issue of what characteristics, besides the basic governing principles of moral behaviour, the machines need in order to be more like human agents. The public opinion about the safety of modern technologies,

2 This type of classification cannot be found in the available body of literature, but is a synthesis of our own examination of the field and corresponding extraction of general questions and noteworthy ongoing lines of research.

in this case, moral machines, is an important aspect of making their usage possible. For that reason, the acceptance of machines as integral parts of society is one of the central themes in Artificial Morality. This problem is probably best verbalized as a question of "what is needed for machines to be *perceived* as moral?".

Lastly, the third part will deal with the technical side of moral engineering. It is necessary to decide the way in which these machines will run, that is, what the best approach for implementing moral algorithms is and what kinds of algorithms should be implemented in the first place. Engineers, in cooperation with psychologists and philosophers, are trying to decide which governing ethical principles or machine learning algorithms will give optimal results in real-life conditions and render correct ethical judgements. Moreover, the choice on a conceptual level of a machine's functioning (whether there is going to be a set of basic principles which govern machine behaviour, or if the machine will be able to learn and extract ethical principles from experience and then use them to guide its own moral judgements) implies a specific programming approach, which, then, has its own technical challenges.

Artificial intelligence has been a growing field of work for the past 50 years (Malle, 2015, pp 161), and yet efforts to answer certain questions about morally functioning autonomous AI machines, or artificial moral agents (hereinafter AMAs), had begun only a decade ago (Yampolskiy, 2013, pp 389). A key reason for even stepping into this endeavour of creating AMAs was the rapid development of autonomous machines or decision-making algorithms used in a wide range of everyday situations, from driverless vehicles and elder care robots, to bank intelligent money transfer software (Wallach & Allen, 2009, pp 17; Goodall, 2014, pp 93, Misselhorn, 2018, pp 162).

Consequently, this emerging usage of autonomous systems has increased the number of situations in which they will be put in a decision-making role with a different magnitude of repercussions for the society. Moreover, there are already seemingly paradigmatic examples of the aftermath of judgements in morally oblivious AIs. There have been incidents in which these AIs, as a result of their reasoning process, selected violent videos for children, produced racist tweets or even racially discriminated against convicts on parole when accessing their risk for recidivism (Shank, DeSanti & Maninger, 2019, pp 652). However, there are even more moral decision-making opportunities that we encounter daily. Although they are not as visible as aforementioned scenarios, and thus not used as representative examples for the exigency argument about the implementation of moral decision-making abilities in artificial intelligence, they are vastly frequent and, consequently, more important: For instance, we can briefly focus on the increase of daily usage of automated vehicles and elder care robots. Goodall addresses (2014) the remark that people rarely make moral decisions while driving, and thus

machines shouldn't either, by accentuating the morality of everyday decisions, regardless of how small they may seem (especially when an evaluation about their importance is made based on actualized consequences rather than possibilities). Accordingly, he states that the category of ethical judgements includes cases such as a driver deciding whether to unlawfully speed up so there can be more room for a cyclist on the road (Goodall, 2014, pp 97). Similarly, Wallach and Allen mention the example of medication dispensing robots for the elderly (Wallach & Allen, 2009, pp 15). In its way of completing the task of handing medicine to someone in need, a robot may encounter various obstacles that require ethical judgement about the robot's further behaviour. What if the mentioned obstacle is a child instead of an object? Would the robot's judgement be based on the utility of alternative solutions? Should the robot have a predefined set of preferable actions and rules it follows, or should it be able to learn from experience and examples of correct judgements in order to abstract guiding rules?

There is a shared concern about the possible outcomes of self-guided behaviour in morally oblivious machines (Anderson & Anderson, 2007, pp 16; Goodall, 2014, pp 94; Misselhorn, 2018, pp 162; Yampolskiy, 2013, pp 389; Shank, DeSanti & Maninger, 2019, pp 649; Wallach & Allen, 2009, pp 3), but also a research field that aims to overcome these concerns. This field is called Artificial Morality. Its central approach to preventing possible judgement mistakes of intelligent machines is ensuring that their behaviour towards humans and the environment is ethically acceptable, which is achieved by creating artificial moral agents, AMAs.

2. Modality of AMAs: moral agency of machines

One of the main problems in Artificial Morality is whether machines can be moral agents in the same way that humans are, or at least moral enough to be attributed the characteristic of moral agency. Following the latter thought, there is a discouraging picture of AI's morality in relation to human morality. The public opinion of AIs is more negative than positive, that is, people are distrustful towards intelligent machines and they do not feel at ease about machines making autonomous decisions. In other words, people do not perceive AIs as moral agents, nor do they attribute to them the status of equal members of the society (Bostrom & Yudkowsky, 2014, pp 318). From such state arises a new problem of inequality amongst humans and AIs. Dispositions that can be formularized and implemented in artificial, intelligent machines, which can then simulate them successfully, often get post hoc characterized as not real enough, or even completely disregarded, because of the idea of non-human embodiment (Bostrom & Yudkowsky, 2014, pp 318). Bostrom states (2014) that this kind of rejection of valuable human characteristics, when they are exhibited by machines, emerges from

the recognition of their specialization in a specific domain. For example, AI's ability to play chess or Go better than the champions in these games ceases to be perceived as extraordinary, impressive or valuable because of the awareness that this ability in AI is limited only to this domain (Bostrom & Yudkowsky, 2014, pp 318). This devaluation of human abilities which are not proven as general traits, but instead exist only for a specific purpose, indicates that value is attributed to those characteristics that are applicable in a variety of situations.

In addition to the demand for generalizability of traits, so they can be accepted as human-like, there is also a demand for a less perfect performance (Indurkhya, 2019, pp 108). Perfection and lack of mistakes in a machine's performance of tasks evokes a sense of mannerism and artificiality in humans. Because of the social rejection of AIs manifestation of human dispositions, the efforts for creating widely accepted machines are going in the direction of making their behaviour more human-like. For example, there have been cases of deliberately constructing AIs that make mistakes while performing specific actions such as dancing or drawing (Indurkhya, 2019, pp 109). This issue of public acceptance and required competence for equal and human-like machines will be addressed in the section *Moral competence of machines*. This section will focus on the conditions of moral agency.

There is no universally accepted definition of moral agency in ethics literature. Furthermore, there are frequent disagreements over what constitutes a moral agent (Misselhorn, 2018, pp 163), but despite this division of opinion, there is also a surprising overlap in different understandings of moral status (Misselhorn, 2018, pp 163).

One of these understandings (Misselhorn, 2018, pp 163) highlights two central conditions of moral agency: (1) the subject must be an *agent*, (2) and it must be a *moral* agent.

Agency is then defined through concepts of self-origination and self-reasoning. The concept of self-origination refers to the origin of an agent's action. The agent is here understood as self-originating only if the source of her action is within herself. That means that the action initiators are the internal structure and dispositions of the subject and not external events. The most demanding form of the self-origination concept refers to "the action without any prior cause" except the agent's humour, but a less strict and commonly used form of self-origination is understood as actions that are under the control of the agent, are not solely determined by external stimuli and can be manifested with "greater flexibility that is dependent on the agent" (Misselhorn, 2018, pp 163). In a practical sense, applicable to artificial agents, less demanding criteria of self-origination means that agents are able to interact with the environment, to affect the environment and its own state without the influence of external events, adapt to external conditions or actively change them. The self-reasoning concept considers the capacity to

act for a reason, in other words, the capacity to have a belief in something and a pro-attitude (desire) towards something. The combination of belief and pro-attitude constitutes a reason to act and guide our behaviour.

Furthermore, moral agency is attributed to the agent if her source of action, and reasons for it, come from inner moral reasons. That is, the agent can be a moral agent only if her self-origination and self-reasoning capacities include moral attributes (Misselhorn, 2018, pp 164).

A similar understanding of moral status (Bostrom & Yudkowsky, 2014, pp 321) also extracts two important criteria: sentience and sapience of the agent. Sentience applies to the ability to have qualia, an idiosyncratic phenomenological experience. Qualia is often understood through the capacity to feel pain, but it refers to any kind of emotional or sensory experience. Moreover, it is thought that animals possess, in different degrees, this ability of phenomenological experience. The concept of sapience is understood as a capacity for self-awareness (consciousness) and acting for a reason. This kind of capacity implies higher cognitive structure which can only be found in humans. It can be noticed that sapience incorporates both Misselhorn's conditions of agency (self-origination and self-reasoning) but does not imply moral reasons that she highlights as necessary for moral agency.

It is clear that artificial systems cannot meet the most demanding forms of aforementioned conditions of moral agency, but given that such metaphysical concepts evoke still unresolved debates concerning human agents, there is a justified reason to concentrate on less demanding criteria of moral agency. In the case of the self-originating concept, we should move away from the metaphysical controversy of determinism and initiation without any prior cause. Less demanding criteria understand self-originating agents as agents who can change their environment or their own state without being influenced by external stimuli (Misselhorn, 2018, pp 163). This criterion puts focus on observable elements of situations that guide our conception and attribution of agency. We argue that this form of conclusion about the agency is a justified way of judging about the moral status of machines, given that it appears to be an important aspect of judging about agency when it comes to humans.

When attributing the cause for someone's behaviour, people primarily take into account the situational factors of the event (Kelley & Michela, 1980). Whether the cause of someone's action is going to be attributed to their internal dispositions or to external situational factors, depends on the observable characteristics of a situation. According to the empirically sustained Kelly's Attribution theory, if there is a possible situational explanation for someone's action, the cause of action will be attributed to the external stimuli rather than person's dispositions (Kelley, 1973; Kelley & Michela, 1980). For example, we do not interpret a professor delivering a lecture as her being a talkative person nor do we interpret a waiter's pleasantness as him being a

friendly person. Instead, we exhibit a tendency to explain their behaviour as situationally structured – the professor talks because it is her job to give a lecture, and the waiter is pleasant because his job also depends on his positive attitude. In lack of congruent situational factors, the cause of action will be attributed to inner factors, a person's dispositions (Kelley & Michela, 1980). That means that, for example, we will interpret the waiter's unpleasantness as him being a rude person because there are no relevant, congruent external factors that can overrule attribution to inner, dispositional factors.

The same framework can be applied to the general attribution of agency. If such regularity of attribution of dispositions is noticed when it comes to human actions, then the same logical line should be justifiably followed by a discussion about the machine's actions. Bostrom's principle of ontogeny non-discrimination (Bostrom & Yudkowsky, 2014, pp 323) also states that "if two beings have the same functionality and the same consciousness experience, and differ only in how they came to existence, then they have the same moral status". That means that if artificial systems can act without any situational factors that noticeably influence their actions, they can be attributed with dispositional causes. These inner causes are markers of agency, and if AI systems have the capacity to act according to inner causes and reasons, they will have some status of agency. Additionally, if those reasons are moral reasons, they will have the status of moral agency (Misselhorn, 2018, pp 164).

An interesting view of agency, applicable to AI systems, is provided within the framework of moral psychology. Gray and colleagues (Gray, Young & Waytz, 2012, pp 103) discuss moral agency (i.e. moral judgement) as fundamentally dependent on, and determined by, mind perception. On the basis of extensive research of mind perception, they conclude that people perceive minds through two independent dimensions – the dimension of *experience* and the dimension of *agency*. The experience dimension is analog to Bostrom's concept of sentience, and is understood as the ability for sensation and feelings, while the agency dimension, which can be represented by the concept of sapience, refers to the capacity to act and to intend (Gray, Young & Waytz, 2012, pp 103). These dimensions of mind perception appear to be strongly linked with perception of one's moral status, usually defined through ascriptions of rights and responsibility. Perception of experience is correlated with ascription of rights, that is, with the perceived ability to feel (pain, pleasantness) comes the ability to benefit or suffer. Perception of agency, on the other hand, is correlated with ascription of responsibility, namely, if one is prescribed a higher capacity to act and intend, one could also be attributed more blame or praise (Gray, Young & Waytz, 2012, pp 104).

As Gray and colleagues define it, perception of agency qualifies moral agents and perception of experience qualifies moral patients (Gray, Young & Waytz, 2012, pp 104). As agency and experience (or moral agency and moral patiency) are independent dimensions, there can be entities high in both

dimensions, low in both dimensions, or high in one and low in the other dimension. For instance, adults are perceived as entities that are high in both agency and patiency, and thus can be both responsible (blamed) for their actions and deserve rights (protection) from actions of others. Moreover, AI systems would be perceived as high in agency, which would grant them the status of moral agents, but low in experience, which would deny them the status of moral patients. Essentially, that means that AI systems will always be perceived as entities who act but never receive (feel). Given the omission of perceived capacity for sensation, AIs will not have moral rights, but, given the actualized perception of agency, they will be ascribed to full spectrum of moral responsibility

Moreover, morality is broadly understood as a dyadic interaction between two perceived minds, a moral agent and a moral patient. Gray and colleagues argue that the essence of morality can be captured in this cognitive template of "perceived intentional moral agent and a suffering moral patient", where the presence of moral agent is required but the presence of suffering moral patient can just be imagined (Gray, Young & Waytz, 2012, pp 107).

A dyadic structure of morality recognizes the phenomena of moral typecasting. Moral typecasting refers to the categorization of people either as moral agents or moral patients. Even though this kind of mutually exclusive categorization is apparent within a specific moral context (where a prototypical moral situation revolves around the interaction of a moral agent and a moral patient), moral typecasting suggests a more general categorization – people are usually and consistently seen as *either* moral agents *or* moral patients (Gray, Young & Waytz, 2012, pp 113).

Furthermore, moral typecasting can influence the perception of one's mind, that is, the perception of one's moral status. Those that are categorized as moral agents are ascribed with the capacity for agency and intention, and are given moral responsibility as well, whereas those categorized as moral patients are ascribed with the capacity for experience and are given moral rights (Gray, Young & Waytz, 2012, pp 113). Given that AIs will consistently be found in roles of moral agents, as acting entities with aims and tasks, they will automatically be categorized as moral agents and correspondingly attributed with agency and intention.

Concluding this section, we can see that AIs, by the very fact of fulfilling the roles of agents, can be (and will be) perceived as *moral agents* with a certain level of expected moral responsibility. However, a natural consequence of typecasting AIs as moral agents will create a general, conclusive perception of them only as moral agents, but never moral patients. This puts AIs in an unflattering position. Although they can have moral agency and can be blamed for their actions, they cannot enjoy the status of being moral patients similar to humans or animals, and will thus not be given corresponding moral rights.

3. Moral competence of machines

The discussion about machine morality has so far been focused on their capacity to be moral agents and the problems of defining moral agency. With those tasks ahead come many difficulties about finding one universally accepted definition of moral agency and choosing which of the many understandings of moral agency to follow when deciding about the machine's moral status. Malle, however, proposes a new approach to the problem (Malle, 2015, pp 245): it is more functional to focus on the constituents of human moral competence and use them as orientation guides for creating morally competent machines, instead of focusing on defining moral agency. Understanding the elements of human moral competence can serve as a guide for the making of moral algorithms for machines. This approach ends discussions about machines as moral agents equivalent to humans, and makes room for more fruitful possibilities for designing machines that are competent agents which can perform the needed tasks. They can also have different degrees of competence.

If machines adequately exhibit this moral competence, people can decide on whether they are willing to accept and form social relationships with the machines. Malle's approach of observation of human behaviour as a guideline for designing machines emphasize the relevance of human-like abilities in AIs. Other authors emphasize this approach as a relevant and successful way for accelerating AIs social acceptance as well (Bostrom & Yudkowsky, 2014, pp 317; Malle, 2015, pp 253; Malhotra, Kotwal, Dalal, 2018, pp 4; Indurkhya, 2019, pp 110).

3.1. Human-like competence in machines

Moral competence is an aptitude to successfully perform moral tasks, namely, tasks of moral decision making and moral behaviour (Malle, 2015, pp 255). Furthermore, moral tasks imply the capability of moral cognition that is defined through one's aptitude for judgements of blame and permissibility, recognition of right and wrong, and emotional reactions while performing these moral tasks (Malle, 2015, pp 255). Acceptance of AIs as moral and social agents depends on their ability to meet people's expectations about their moral and social responsibility. The initial idea is that, with performing regular human tasks, AIs will also take on regular human responsibilities. Their capacity to satisfy these expectations, and successfully perform moral tasks, determines in what degree they are perceived as equal members of society (Bostrom & Yudkowsky, 2014, pp 316; Malle, 2015, pp 245).

Central elements of human moral competence, according to Malle, are (1) moral vocabulary, (2) a system of norms, (3) moral cognition and affect, (4) moral decision making and action, (5) and moral communication (Malle, 2015, pp 245). An extensive study of these elements can be found in Malle,

2015, but the highlighting of the importance for machines to demonstrate more human-like characteristics, in order to make them optimal social agents, puts focus on the emotional aspect of human functioning, that is, on the needed emotional aspect of machine functioning.

Moral philosophy and moral psychology dominantly concentrated their research of morality around the study of moral reasoning, thus neglecting moral emotions, up until the 1990s. This leadership of cognitive reasoning in understanding morality was a product of cognitive revolution and the idea that morality, like language, can be expressed through underlying cognitive structures and corresponding transformations (Haidt, 2003, pp 852). Later theories, on the other hand, highlighted the role of emotions, but the most realistic approach to this problem is the comprehension of both moral cognition and moral emotions as backbones of human morality.

The capacity for both moral cognition and moral emotions that humans exhibit lacks in the case of AIs. As discussed in the previous section, machines can be understood as moral agents with an expected moral responsibility, but never as moral patients with related moral rights. AIs are presumably denied moral patiency because they are missing the capacity for qualia. This capacity, besides sensory experiences such as pain, incorporates an emotional life of an entity, that is, a potential for emotional experience. Emotions, or emotional experiences, are reactions to inter– and intra-activity of an organism, with the main function of mobilizing that organism to adaptively deal with such encounters (Ekman, 1999, pp 46). In other words, emotions are mainly responses to threatening and beneficial stimuli with great motivational tendency, attendant facial expressions and phenomenological experience (Haidt, 2003, 854).

The difference between emotions and moral emotions lies in their relation to self (Haidt, 2003, pp 853). According to Haidt, moral emotions are those emotions that are not directed to self but are "linked to the interests or welfare of other people or a society as a whole", whereas other non-moral emotions are always in more direct relation to self and occur as a reaction to influences on the agent. AIs are missing both types of emotional experience. Emotions such as fear, sadness and happiness are mainly categorized as non-moral emotions, given their occurrence in situations directly related to the agent or in situations of less direct relation between the self and the other. Lack of these emotions deprives AIs of moral rights because not only can they not be physically hurt but they are not able to feel emotional pain or gain either, and are thus perceived as entities that do not need to be protected by society, i.e. do not need moral rights. Moreover, the most prototypical moral emotions are elevation, anger, guilt and compassion, as their triggers are usually disinterested stimuli and are easily triggered by tragedies and transgressions of strangers (Haidt, 2003, pp 854). AI's inability to feel guilt if it makes a judgement error and causes tragedies, or to feel compassion or anger

if it encounters tragedy and pain, determines its further behaviour. Given that emotions have strong action tendencies and motivate some kind of response to the eliciting stimuli, AI's emotional oblivion restricts its empathic and helping actions. That influences the social perception of machines' "coldness" and elicits anticipation of their reluctance to help, which again accelerates people's distrust in machines and makes their social acceptance difficult.

Even though machines can be implemented with algorithms of moral acting, and can thus help others and intervene in situations of need, they are still perceived as agents that cannot feel the direct consequences of moral behaviour related to them. Such picture of senseless entities restricts the attribution of moral patiency and makes them humanly distant.

There are, however, other traits that will help AIs to be socially accepted. Bostrom adds several criteria that need to be algorithmically formalized and implemented in machines (Bostrom & Yudkowsky, 2014, pp 317). The central one, to which others may be reduced, is transparency in decision making. The transparency of AIs reasoning process enables its inspection, a matter of significant importance in the possible scenarios of reasoning mistakes or hazards caused by AIs decisions (Bostrom & Yudkowsky, 2014, pp 317). The knowledge of how these intelligent algorithms make their decisions does not only enable the tracking of responsibility (and blame) of machines but also has the purpose of amplifying their social trustworthiness. This openness to investigation removes their "black box" artificial invisibility and excites their similarity to human behaviour. Therefore, it is more than needed to equip AIs with psychologically relevant explanations of their own processes (Indurkhya, 2019, pp 110).

Being equipped with psychologically compelling explanations, such as transparency of processes, may also excite AIs general similarity to human behaviour and their consequential acceptance. Until the wanted level of technical development is reached, and AIs are endowed with senses, further development of machines needs to progress in the direction of psychological openness of their judging processes.

3.2. Responsible AIs

An individual's involvement in society is in social psychology often discussed from the perspective of interactionism. The same perspective can be applied to machines, given the effort put into making them welcome members of society that have the status of moral agents. Interactionism describes identity as a meaning derived from social roles one occupies (Burke & Tully, 1977, pp 883). Social surrounding reacts to the agent with expectations for the agent's behaviour to correspond with her social role, in other words, social surrounding reacts *as if* the agent's identity is appropriate to her role performance. An agent understands that reaction and forms a meaning about her identity that guides her following behaviour (Burke &

Tully, 1977, pp 883). In the case of AI systems, it is important to highlight that these social expectations are derived from the very fact that someone is a social agent (Stouten, DeCremer & Van Dijk, 2006, pp 894).

When AI occupies a certain social role, it will evoke corresponding social expectations about its behaviour and dispositions that are common in humans (Gray, Young & Waytz, 2012, pp 113). AI systems, in this case, need to prove their identity of moral agents by adequately dealing with expected moral tasks. Moral competence, besides moral judgement and emotions, entails conforming to social norms such as the principles of righteousness and equality. If social expectations of honouring these principles are disappointed, people will react with anger, emotional distress and retributive reactions in order to correct the inflicted injustice (Guth, Schmittberger & Schwarze, 1982, pp 384; Stouten, DeCremer & Van Dijk, 2006, pp 895). Because these reactions only appear when someone is perceived as a moral agent, that is, if someone is perceived responsible for their actions and obliged to follow social norms (Gray, Young & Waytz, 2012, pp 113), we can test the moral status of AIs by examining people's reactions to AIs in situations following the violation of social norms.

There has been a new body of experimental literature that grips the above-mentioned problems of AIs' social acceptance. One of the ways to investigate their social status, or at least to scratch the surface of social interaction between humans and machines, is through the Game Theory experiments. These experiments simulate decision-making interactions between players with an aim to reveal and understand the components of their reactions and reasoning (Osburne, 2004, pp 1). Simulated situations of choice often require social choices where subjects can demonstrate their social norms compliance or violation, and reaction to the compliance or violation of others. That is, they can demonstrate their moral competency.

Relevant for these purposes is the bargaining game, called the Ultimatum Game. The simplest and most commonly used version of the Ultimatum Game is the two-player version. This is a bargaining game because one of the players must solve a distribution problem, usually of goods (Guth, Schmittberger & Schwarze, 1982, pp 367). When this player (commonly known as the first player) makes her choice of distribution, she restricts all the possible alternatives of distribution of goods to one proposal (her choice). The other player (the second player) can then only accept or refuse the first player's proposal. In other words, the first player decides on how to distribute the goods (e.g. money) and makes her proposal to the second player who can then only accept or decline. There are no simultaneous moves of players in the Ultimatum Game, but instead, every aspect of the game is successive so that the players can always observe each other's decisions (Guth, Schmittberger & Schwarze, 1982, pp 367; Osburne, 2004, pp 179). That way, every player is, at the same time, always and completely informed of every previous move in the game[3].

3 Such a game is said to have perfect information.

The specificity of the Ultimatum Game is that the bargaining comes in a form of "strategic reactions based on anticipated future events" (Guth, Schmittberger & Schwarze, 1982, pp 368) where the first player takes into account the "fairness" of her proposal to the second player, and the second player takes into account that the alternative option to the first player's proposal, however unbeneficial, is nothing (and is always worse than the proposed distribution). Because of this bargaining aspect, the game is suitable for investigating social norms and moral behaviour. A further variation of the game can be found in the Dictator's Game with perfect information. The Dictator's Game has only one move in which the first player makes a proposal, and the second player has no other option but to accept it.

The question of interest, when it comes to the social status of machines, is whether they will get the same treatment as human players. Given that both the Ultimatum and the Dictator's games are widely used in social interaction researches, there are noticed regularities of choices that people make and emotional reactions to those choices. If these regularities of human behaviour towards each other also manifest in the games with human and machine players, that is, if human players treat machines the same as they treat humans, there can be a more optimistic comprehension of the machines' social status.

One of the robust findings are acts of retribution when one player feels that norms have been deliberately violated by the other player (Guth, Schmittberger & Schwarze, 1982, pp 384; Stouten, DeCremer & Van Dijk, 2006, pp 895). In situations where the first player's proposal exceeds the 70:30 proportion of distributed goods in her favour, the second player would usually decline the offer even though it means that she will end up without anything. This kind of reaction is described as a retributive reaction to what someone understands as injustice (Stouten, DeCremer & Van Dijk, 2006, pp 895). There is also an observed regularity of the prosocial proposals first players commonly make. In most cases their distributions are fairly made, that is, the majority of players distribute goods in an approximately equal share. More specifically, they strive to benefit from their distribution, but to also split the goods according to the fairness norm (Forsythe, Horowitz, Savin & Sefton, 1994, pp 362, Guth, Schmittberger & Schwarze, 1982, pp 384).

Equivalent treatment of machines and humans was demonstrated in one such experiment (Nagataki et al., 2019). No significant difference was found between human and robot status in prosocial and retributive tendencies of participants. All human participants made the same prosocial offer of nearly half of the total amount of money to robots as they did to humans, in both, the Ultimatum and the Dictator's Game. That way, they equally respected the norm of fairness amongst social agents, whether the other agent was human or not. Moreover, the participants rejected "unfair" offers from robots, just as they did from humans, thus demonstrating a will to punish what they

considered unjust behaviour, even at the cost of their own gain. These kinds of equivalent reactions to machines and humans may speak in favour of potentially equal social status between them.

4. Engineering approaches to machine morality

Because the consequences of AIs decisions have an unavoidable impact on humans they need to be treated at least as agents with moral behaviour, regardless of the society's acceptance and the question of their full moral agency (Allen, Smith & Wallach, 2005, pp 149). The very idea behind artificial moral agents (AMAs) is to implement human-like characteristics and learning abilities in them so that they can regulate and monitor their own behaviour, correct themselves and perform better in the future decision-making situations (Wallach & Allen, 2009, pp 15). This is the intersection of work paths of engineers, philosophers, and moral psychologists.

Top-down and bottom-up approaches are two traditional engineering approaches that dictate how different moral principles can be used and algorithmically formalized with the goal of creating AMAs. The third, hybrid approach emerges as a combination of the former two and is insofar the most promising one (Wallach, Allen & Smith, 2007, pp 575: Misselhorn, 2018, pp 166).

4.1. Top-down systems

Top-down approaches are based on fixed normative principles, implemented in AMAs, which are then used as guiding rules of the machine's behaviour. Often called a "rule-based" approach (Allen, Smith & Wallach, 2005, pp 150), top-down models require a general set of moral principles that need to be selected. These principles are then universally obeyed in every situation of moral dilemma and expressed throughout the machine's actions.

One of the first problems with these systems is the selection of moral principles to begin with. There can be an unlimited set of contents from which these main principles can be selected (Allen, Smith & Wallach, 2005, pp 150). Most commonly used moral norms are derived from great ethical theories such as Kantian deontology and utilitarianism, but other frequently named principles are Asimov's laws of robotics, The Ten Commandments or the Torah Commandments (Goodall, 2014, pp 98; Allen, Smith & Wallach, 2005, pp 150; Misselhorn, 2018, pp 166; Yamapolskiy, 2013, pp 389). As it can be seen, these principles can vary dramatically in their generality and number, from three general and unspecified principles in the case of Asimov's laws to the complex computational system needed when it comes to utilitarianism.

Because of their generality, lack of applicability to more domain-specific contexts, and inability to define a concrete set of principles or actions which will guide one's decisions across different contexts and situations, top-down

approaches are severely criticized (Allen, Smith & Wallach, 2005, pp 150). The challenge lies in finding an optimal way of deriving a set of specific rules from the abstract principles. Even though their number is fixed, these rules should be usable in a variety of specific situations.

This approach predominantly uses consequentialist and Kantian theories as starting points in deriving guiding principles for AMAs. Both theories have their own specific problems, but also have a shared one (Allen, Smith & Wallach, 2005, pp 151). The top-down approach based on Kantian deontology encounters the problem of hierarchy of principles, that is, how to submit all its specific principles to one highest principle without contradiction. The other main problem concerns the availability of information, more closely, how AMA should know about the intentions and motives of every agent included in some decision-making situation. Utilitarian AMA faces problems of finding a common value scale for measuring different utilities in various situations and of enormous computational resources needed for even evaluating possible outcomes for every event (Allen, Smith & Wallach, 2005, pp 151). Their shared problem, and a reason for the abandonment of the top-down approach, is the unlikeliness that these algorithms could ever collect and compare every information that they need. This is even more transparent in the cases of consideration of future consequences of actions, instead of focusing on direct and present consequences (Allen, Smith & Wallach, 2005, pp 151).

4.2. Bottom-up systems

The question that imposes itself is how humans restrict their own calculation of continuous external stimuli and predict future consequences since this problem of computational and informational overload is present in the case of their cognitive system as well. Human behaviour is often guided by heuristics and affects decision making (Allen, Smith & Wallach, 2005, pp 151). Moreover, we have the ability to learn from experience and observation. That leads to creating cognitive schemes (scenarios) of plausible events that guide our behaviour when we end up in similar situations (Greene, 2017, pp 69) and, in most cases of decision making, it is what we rely on.

Bottom-up models are based on the abovementioned history of learning, more closely on real data, the experience of correct judgements in decision-making situations from which AMA abstracts moral principles and controls its acts. Bottom-up AMAs do not need an initial set of guiding principles. That means that AMA learns proper moral behaviour while actively participating in their environment (Allen, Smith & Wallach, 2005, pp 151). Bottom-up AMAs can be realized throughout different initial settings and algorithms that determine the type of their learning process. They can simulate learning through trial and error attempts, they can be based on educational learning processes and simulate socialization and the growth of a child, they can simulate evolutionary processes of cognitive and moral growth of an agent,

or they can be based on neural-network processes which associate patterns in the surroundings they encounter (Misselhorn, 2018, pp 166; Wallach, Allen & Smith, 2007, pp 570).

This approach resolves some of the problems that top-down models have by introducing a self-changing and self-improving system. The machines based on Bayesian models can adopt moral rules and change their behaviour when in contact with their social surroundings. They constantly revaluate first guiding principles, as a reaction to new information and experience learning, and verify the consistency of all previously formed rules (Shaw, Stockel, Orr, Lidbetter & Cohen, 2018, pp 73). These machines become self-checking agents capable of human-like adaptation to surroundings.

Bottom-up systems provide more natural and stronger models of moral reasoning that can be an almost ideal approach for creating agents with optimal social functioning and ethically responsible judgements. However, a significant problem of these systems is that they are extremely difficult to develop and usually need a lot of time to evolve into an optimal moral-reasoning autonomous agent (Allen, Smith & Wallach, 2005, pp 151). There is a rising problem of controlling the learning data for AMAs, so cases in which bad data may contribute to their socially unacceptable principles and decisions can be avoided. Other than that, engineers encounter an additional problem of not knowing which principles to use as a guideline in the situations of changed contexts (Wallach, Allen & Smith, 2007, pp 572) and the uncertainty of what will be the evolutionary outcome of a specific AMA (Misselhorn, 2018, pp 167).

4.3. Hybrid systems

Although top-down and bottom-up models represent the most common way of implementation of moral competence in AMAs, their combination is often characterized as necessary for overcoming the specific and general problems that both approaches carry (Misselhorn, 2018, pp 166). Therefore, hybrid systems originate from combining top-down and bottom-up approaches into one "eclectic" model. Hybrid AMAs are implemented with algorithms inspired by both traditional approaches.

Their top-down part is a predefined and fixed set of initial principles that serves as a starting point from which AMAs learn and self-improve. The predefined sets of rules are often not as general as in traditional top-down systems, but are more closely specified to domains in which they are set to be used (Misselhorn, 2018, pp 167). As it was mentioned, hybrid AMAs maintain the ability of self-improvement regardless of their initial moral principles. This ability to learn from experience and adapt is their bottom-up part (Allen, Smith, Wallach, 2005, pp 153), and because of that, their guiding principles often get changed throughout this learning process. Hybrid AMAs, as self-checking agents that are actively involved in the environment, develop

even more specific moral judgements congruent to characteristics of their surroundings (Misselhorn, 2018, 166).

Allen, Wallach and Smith (2005) interpret top-down and bottom-up systems through contrasting the explicit and implicit values and their ways of acquiring. In their description, top-down systems can be understood as explicit values and ethical principles "outside of the entity" that are demanded from a specific cultural milieu, while bottom-up systems are implicit moral values abstracted from practice and experience that then emerge from "within the entity" (Allen, Smith, Wallach, 2005, pp 153). That way, a top-down AMA can be described as AMA of "rights and duties" or "welfare and utility", while bottom-up AMA is an AMA of "practice and experience". As their combination, hybrid AMA is understood as an entity raised in a culture which prescribes its own explicit moral concerns and judgements and requires they be respected, while it (the AMA) still has constant opportunity to discover and learn other values and traits from practice (Wallach, Allen & Smith, 2007, pp 576). That is, AMA is given some kind of parental rules (like those a child is demanded to follow during his or her development) but it also interacts with the environment and through that learns or demonstrates her individual traits.

Because of the above-described hybrid AMA's position between top-down and bottom-up models, Aristotelian virtue ethics is seen as a fruitful framework for hybrid algorithms (Wallach, Allen & Smith, 2007, pp 576). Aristotelian virtuous character resembles ethical principles and initial rules implemented in hybrid AMA because, in both cases, they are initial motivators of one's action and overall behaviour, and are understood as one's features or traits. Furthermore, Aristotelian understanding that moral virtues are learned through experience, and consolidated from constant practice, is equivalent to hybrid AMA's ability to learn from experience and self-improve (Wallach, Allen & Smith, 2007, pp 576).

Hybrid AMA begins to be interpreted as an individual, into which one can instill Aristotelian "good traits", virtues or characteristics, "complex patterns of motivation" (Wallach, Allen & Smith, 2007, pp 577) and personality dispositions that determine our tendencies to act in a certain way. This new line of thinking has the reincarnated problems of top-down models. It is difficult to choose a number and type of virtues machines should have in order to be moral, but importantly, it is extremely difficult to simulate virtues (Wallach, Allen & Smith, 2007, pp 577). Virtues, as dispositions and patterns of motivation, are manifested in one's general way of behaving, i.e. in a variety of different situations. In that sense, one virtue has multiple behavioural expressions and is responsible for various acts. Because of that, AMAs should be able to connect every potential action or judgement to a certain trait, that is, they should be able to "know "every possible manifestation of some virtue so it could adequately practice that virtue in its overall behaviour. Moreover, the traditional problem of constant checking if every chosen action is

congruent with all guiding principles, both specific and higher, requires enormous computational power. Even more computer power is needed for creating a non-contradictory hierarchy of virtues and enabling a changeable AMA to develop and incorporate new virtues in such a non-contradictory way (Wallach, Allen & Smith, 2007, pp 577).

Another approach of implementing Aristotelian virtues in AMAs comes from bottom-up models, specifically, neural networks. The central idea is the development of a virtuous character. Neural network system has access to training data from which it abstracts moral principles, while the further gathering of data is realized in real-life scenarios where network surpasses its previous generalized principles (Wallach, Allen & Smith, 2007, pp 577). However, present perspectives that provide insights into human developmental process still cannot provide adequate frameworks for the learning process of moral virtues when it comes to neural networks (Wallach, Allen & Smith, 2007, pp 578), and for now, this approach remains only a daring idea.

4.3.1. Culturally assimilated AMAs

Allan et al. (2005) discuss the initial set of guiding principles in hybrid AMAs as explicit values of the cultural milieu they are made for. However, this thought is not just a superficial analogy made for better conceptual understanding of the top-down approach. Cultural variation requires the adaptation of machines to specific contexts in which they function. For this to be done, we need to first abstract specific dimensions of morality, and from there conclude which specific dimensions suit which culture.

The Moral Foundation Theory (MFT) provides a conclusive picture of a moral mind "constructed" of a universal set of moral modules, innate foundations which guide the learning process of moral values, norms and rules, and are environmentally sensitive (Graham et al., 2013, pp 10). Haidt and colleges integrate the evolutionary position of innate morality and a constructivist perspective on cultural shaping of values and moral behaviour. They propose that the human moral mind is organized "in advance of experience", that is, it evolved a set of "moral matrices" (shared knowledge) as a tool for solving social problems of a cultural human (Graham et al., 2013, pp 8). These modules are understood as foundational moral instincts that enable the learning of some moral values and behaviours over the other. That way, people are innate with potential for acquiring a set of universal moral norms (foundations). Which of these universal norms will be adopted, which particular values generated and in what degree, will be determined by a specific culture, through one's development process.

Based on the MFT perspective on moral norms we can further discuss which set of initial guiding rules, or moral values, should be implemented in AMAs. MFT proposes five moral foundations that we mark as suitable for AMAs' norms (Graham et al., 2013, pp 12).

(1) The Care/harm foundation represents a functional mechanism that enables association of perceived suffering with actions of nurturing, caring and protection. This foundation is extremely important for machines with highly responsible tasks that revolve around people, such as elderly or children care, but also for machines whose judgement decisions may directly or indirectly influence one's life (automated vehicles). The AMA with values of caring for– and protecting others will presumably be of equal importance across cultures, given that more trustworthy and reliable machines will be valued and demanded– regardless of the individual differences between individualistic and collectivist cultures.

(2) The Fairness/cheating foundation is responsible for being observant to sings of cooperation or cheating amongst others (Graham et al., 2013, pp 13). It generates specific values such as righteousness, fairness, sensitivity to inequality, justice that include retributive behaviour as well as rewarding acts, and so on. Dimensions of Care, Fairness and Sanctity (described in the next paragraph) turn out to be important categories for evaluation of virtuousness in both liberal and conservative groups (Graham et al., 2013, pp 20). Given this invariability to conservativism, the value of fairness would be important for machines to have in different cultures or social groups they are made for. By satisfying social demands for fair judgement, AMAs would prove themselves as responsible and trustworthy members of society.

(3) The Sanctity/degradation foundation relates to sensitivity for puritanism of body and "soul", that is, values and motives for which "people treat their bodies as temples" (Graham et al., 2013, pp 14). It is, as mentioned above, a valuable moral norm for estimation of virtuous character, but it is not invariant to cultural context. The sanctity is extremely important in collectivist and traditional cultures, where AMAs need to adapt to bigger roles of purity and religious concerns in everyday life (Graham et al., 2013, pp 26). That requires implementation of religious beliefs congruent to the market culture of an AMA. Just as it would be preferred that AMAs exhibit dominantly practiced and expressed religious rules, values and norms in Eastern cultures, it would be required that AMAs do not exhibit those same values in a secular society. Moreover, intragroup differences in cultural variation are robustly greater than intergroup differences, that is, these traditional differences are greater within cultures than between themselves (Graham et al., 2013, pp 26). Given regularities such as this, it is better to equip AMAs with values of puritanism and religion according to the tasks they will perform. If the deciding process of the task requires evaluation of such criteria, then its implementation is also needed. These are not just complex tasks, but tasks for which optimal solutions involve cultural knowledge. We already mentioned the example of medication dispensing robots for the elderly (Wallach & Allen, 2009, pp 15). In its way of handing the medicine, a robot may encounter various obstacles that require judgements about whether they are religious objects that need to be carefully avoided or not.

(4) The Loyalty/betrayal foundation highlights the importance of motivational tendencies to exhibit the traits such as agreeableness, fidelity and alliance, because of their significance for forming coalitions and preserving group cohesion (Graham et al., 2013, pp 13). Compliance to this norm makes social functioning, particularly group functioning, possible and thus is an inevitable value for AMAs that are privately owned. Even though loyalty foundation is more connected to the conservative groups (Graham et al., 2013, pp 16), machines need to exhibit alliance tendencies as an acceptance tool.

(5) The Authority/subversion foundation serves as a mechanism for navigating one's behaviour in hierarchical social interactions. It shapes values of obedience and deference (Graham et al., 2013, pp 13) that are, again, more valuable in conservative groups and collectivist cultures than in liberal and more individualistic groups. These values may suit AMAs who have roles of carers and are in direct contact with humans.

All five moral foundations interact with the environment and generate more specific moral values. These foundations are thought to be universal structures of the human moral mind, but their shaping and development is vastly dependent on culture. As we can see, less traditional and liberal groups generate the care/harm and fairness/cheating foundation in greater degree than conservative groups, and more traditional and a conservative environment values the authority, loyalty and sanctity foundations more than liberal groups (Graham et al., 2013, pp 16). That does not exclude some moral norms from certain cultures, but rather priorities values within cultures. Creators of AMAs should thus be sensitive to these cultural moral priorities when making machines for targeted markets, and MFT provides an inclusive and culturally sensitive framework for this kind of deliberation of initial guiding values.

5. Conclusion

This paper had the aim of systemizing the complex, even though new and yet expanding, field of Artificial Morality. Artificial Morality centres around the idea of artificial moral agents (AMAs) which represent self-checking machines able to change and grow while making moral decisions side-by-side with humans. The presented structure of main problems in Artificial Morality originated from the authors themselves. These problems, even though noticeable research obstacles, have never, to our knowledge, been understood as a set of three conceptual problems – philosophical, psychological and a technical one.

In the beginning, we had to inspect the question of moral agency and its theoretical applicability to machines (a philosophical problem). Hopefully, we have given our own insights by proposing a line of thinking about machine

agency similar to the understanding of human agency. In accordance with the less demanding frame of agency, if artificial systems can act without any situational factors that noticeably influence their actions, they can be attributed with dispositional causes just as humans do. Moreover, if AIs exhibit the capacity to act based on their inner causes, dispositions, they will have the agency status. This agency status can be understood as the status of moral agents if those inner causes and reasons were moral reasons.

The problem of social perception and acceptance of AMAs (psychological problem) has a potentially optimistic solution. It was shown that people react to machines in the same way as they do to humans, thus attributing to them social status and responsibility for their actions. Given that the empirical literature on this subject is still limited, these results should be taken with caution, and used more as an implication for further research rather than conclusions. While considering the importance of human-like competence for acceptance of AMAs, a new line of research has emerged. A great deal of effort has been invested in implementing some additional instances such as emotions, consciousness, or other human capacities (like the theory of mind and symbolic understanding), as it is believed that only these competencies can make reliable and fully moral artificial agents. These competencies are indispensable parts of AMAs, not only because of their social acceptance but for their better functionality (Allen, Smith, Wallach, 2005, pp 153). This problem has not been inspected in detail, but its significance for creating functional AMAs will determine following research in the field.

In the end, we have given a brief overlook of the current state of technical advances, possibilities and restrictions in developing a fully functional AMA (technical problem). There are three main approaches to implementation of moral capacities in machines: the top-down, bottom-up and hybrid approach. The first two approaches are traditional and most commonly used systems that are being gradually replaced by their combination, a hybrid model, as they provide only partially functional AMAs. However, hybrid systems prove to be out of the current theoretical and technical reach. Existing frames of learning processes of human moral competence are still incompatible with the mode of neural networks which are integral to hybrid and bottom-up systems. These inspiring ideas in machine learning, even though challenging endeavours, will also determine future efforts in creating AMAs.

We conclude this section with discussion on culturally sensitive hybrid AMAs. From the perspective of the Moral Foundation Theory (Graham et al., 2013), we suggest five moral norms that should be closely evaluated when deciding which initial guiding principles should be implemented in machines. Moreover, we draft some guidelines for acknowledging cultural differences in the valuation of such moral norms but do not offer final solutions. The field of moral psychology, particularly the study of universal moral rules and cultural variation in norms and practices, is still in its developing research stage, and until we have a clearer picture of human morality there will be a limited potential for this kind of extrapolation onto machines.

References

Allen, C., Smit, I., & Wallach, W. (2005). "Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches". *Ethics and Information Technology*, 7(3), 149–155.

Anderson, Michael & Anderson, Susan. (2007). "Machine Ethics: Creating an Ethical Intelligent Agent". *Ai Magazine*. 28. 15–26.

Bostrom, N. and Yudkowsky, E. (2014) "The Ethics of Artificial Intelligence". In: Frankish, K. and Ramsey, W., Eds., *Cambridge Handbook of Artificial Intelligence*, Cambridge University Press, New York, 316–334.

Burke, P. J., & Tully, J. C. (1977). "The Measurement of Role Identity". *Social Forces*, 55(4), 881–897.

Ekman, P. (1999). "Basic emotions". *Handbook of cognition and emotion*, 98(45–60), 16.

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). "Fairness in simple bargaining experiments". *Games and Economic behavior*, 6(3), 347–369.

Goodall, N. J. (2014). "Machine Ethics and Automated Vehicles". In *Road vehicle automation.* Springer, Cham, 93–102.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). "Moral foundations theory: The pragmatic validity of moral pluralism". In *Advances in experimental social psychology* (Vol. 47, pp. 55–130). Academic Press.

Gray, K., Young, L., & Waytz, A. (2012). "Mind perception is the essence of morality". *Psychological inquiry*, 23(2), 101–124.

Greene, J. D. (2017). "The rat-a-gorical imperative: Moral intuition and the limits of affective learning". *Cognition*, 167, 66–77.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). "An experimental analysis of ultimatum bargaining". *Journal of Economic Behavior & Organization*, 3(4), 367–388.

Haidt, J. (2003). "The moral emotions". *Handbook of affective sciences*, 11(2003), 852–870.

Indurkhya, B. (2019). "Is morality the last frontier for machines?". *New Ideas in Psychology*, 54, 107–111.

Kelley, H. H. (1973). "The processes of causal attribution". *American Psychologist*, 28(2), 107–128.

Kelley, H. H., & Michela, J. L. (1980). "Attribution Theory and Research". *Annual Review of Psychology*, 31(1), 457–501.

Malhotra, C., Kotwal, V., & Dalal, S. (2018, November). "Ethical Framework for Machine Learning". In 2018 ITU *Kaleidoscope: Machine Learning for a 5G Future (ITU K)* (pp. 1–8). IEEE.

Malle, B. F. (2015). "Integrating robot ethics and machine morality: the study and design of moral competence in robots". *Ethics and Information Technology*, 18(4), 243–256.

Misselhorn, C. (2018). "Artificial Morality. Concepts, Issues and Challenges". *Society*, 55(2), 161–169.

Nagataki, S., Ohira, H., Kashiwabata, T., Konno, T., Hashimoto, T., Miura, T., ... & Kubota, S. I. (2019, June). "Can Morality Be Ascribed to Robot?". In *Proceedings of the XX International Conference on Human Computer Interaction* (p. 44). ACM.

Osborne, M. J. (2004). *An introduction to game theory* (Vol. 3, No. 3). New York: Oxford university press.

Shank, D. B., DeSanti, A., & Maninger, T. (2019). "When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions". *Information, Communication & Society*, 22(5), 648–663.

Shaw, N. P., Stöckel, A., Orr, R. W., Lidbetter, T. F., & Cohen, R. (2018, March). "Towards provably moral AI agents in bottom-up learning frameworks". In *2018 AAAI Spring Symposium Series*.

Stouten, J., De Cremer, D., & van Dijk, E. (2006). "Violating Equality in Social Dilemmas: Emotional and Retributive Reactions as a Function of Trust, Attribution, and Honesty". *Personality and Social Psychology Bulletin*, 32(7), 894–906.

Wallach, W., Allen, C., & Smit, I. (2008). "Machine morality: bottom-up and top-down approaches for modelling human moral faculties". *Ai & Society*, 22(4), 565–582.

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press.

Yampolskiy, R. V. (2013). "Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach". *Philosophy and Theory of Artificial Intelligence*, 389–396.

Christian B. Miller
Wake Forest University
millerc@wfu.edu

Original Scientific Paper
UDC 17.022.1
171/172
179.9

THE VIRTUE OF HONESTY, NAZIS AT THE DOOR, AND HUCK FINN CASES

Abstract: *I begin by outlining some of the central conceptual features of the virtue of honesty. But the real focus of the paper is on seeing how my account of honesty can handle certain challenging cases. One case is the "Nazi at the door" example. The other is Mark Twain's Huck Finn, who seemed to think that what he was doing in helping Jim was morally wrong, and yet we would be reticent to count it as a case of failing to be honest. I argue that my proposed account of honesty can recommend plausible ways to think about both of these famous cases.*

Keywords: *Honesty, Huck Finn, Lying, Cheating, Stealing, Virtue, Character, Misleading.*

In recent years, philosophers have had plenty to say about certain virtues. Modesty is one of them. Compassion is another. Self-control is a third.[1]

Honesty? Not so much. Indeed, as far as I am aware, only two papers on honesty have appeared in philosophy journals in the last 40 years.[2]

My current research is trying to remedy this situation. In several papers, an edited volume, and a book manuscript, I am trying to direct philosopher's attention to this stunningly neglected virtue and many of the very interesting issues that arise in discussing it.[3]

This paper gets to the heart of this project by outlining an account of the central conceptual features of the virtue of honesty. That happens in section

1 For an example of each, see Driver 1989, Caouette and Price 2018, and Mele 1995, respectively.

2 Smith 2003 and Wilson 2018. There are also a few brief discussions of honesty in monographs, such as MacIntyre 1981.

To some extent the lack of attention to honesty is also found in other fields such as personality psychology, although there has been more work done recently in positive psychology and also using the HEXACO personality trait framework.

3 See Miller 2017, in progress, and Miller and West forthcoming. Some of the interesting issues arise with respect to other virtues as well, but many are specific to the virtue of honesty, such as whether it has a vice of excess, and whether a pluralist theory of motivation is the best way to go.

 DOI: 10.5937/BPA1932051M

one. The remainder of the paper is devoted to seeing how the account can handle certain challenging cases. Section two focuses on "Nazi at the door" examples. Section three turns to cases like Huck Finn, who seemed to think that what he was doing in helping Jim was morally wrong, and yet we would be reticent to count it as a case of lack of honesty. I ultimately conclude that my proposed account of honesty can recommend plausible ways to think about both of these famous cases.

1. The Account of Honesty

The central notion at work in my account is *reliably not intentionally distorting the facts*. An honest person, in other words, is someone who is disposed to reliably not intentionally distort the facts, both to herself or to other people.

I do not have an analysis to offer of "distorting" in this context, and I am doubtful that an informative and reductive account could even be given. But a close synonym to what I have in mind is captured by "misrepresentation" – an honest person is someone who is disposed to not intentionally misrepresent the facts, either to herself or to other people.[4]

I can say more about "intentionally." In some contexts, "intentionally" could be understood as "intending" or "as a result of an intention."[5] The relevant contrast would be "not as part of a plan." So the proposal would amount to saying that an honest person reliably does not plan to distort the facts.

But that is *not* the sense of "intentionally" I have in mind. My usage includes planning and acting from an intention, but it is broader since it includes other mental states as well, such as wants, wishes, emotions, feelings, and the like. Actions results from these mental states need not always be done as part of a plan. Furthermore, "intentional action" on my usage can be causally influenced by subconscious mental states.

So the relevant contrast to "intentionally," as I am using the term here, is "accidentally." If by accident a banana happens to fall into someone's pocketbook at the grocery store and she walks out without paying, she is not intentionally distorting the facts in my sense. Similarly while playing a board game, if no one notices that the players went out of order in a way that benefited one of the them, then that also would not be intentionally distorting the facts. On the flip side, cheating or stealing that is influenced by unconscious psychological states could count as intentional, and so could also count as dishonest in my sense. This would be the case, even if the person

---

4 For a similar claim, see Guenin 2005: 222.

5 For discussion see Bratman 1987. I am grateful to Walter Sinnott-Armstrong for helpful discussion.

didn't recognize those states consciously or they didn't play a conscious role in her planning.

What about "reliably"? Well, clearly someone who refrains from intentionally distorting the facts just once or twice out of hundreds of opportunities does not count as honest. "Reliably" is the label I will use for both stability over time and consistency across situations. Stably over time, the honest lawyer does not intentionally distort the facts in the courtroom. But if she is an honest *person*, and not just an honest *lawyer*, then she typically refrains from distorting the facts at home, at the gym, at the stores, and so forth. To be fair, honesty comes in degrees, and she does not have to be *perfectly* reliable in order to count as honest. Which is good news, as otherwise the virtue of honesty would be incredibly difficult to instantiate.

With these clarifications in mind, I suggest we consider the following as an initial proposal for characterizing the virtue of honesty:

(H1) The virtue of honesty is, centrally, a character trait concerned with reliably not intentionally distorting the facts.[6]

Note that (H1) does not purport to offer necessary and sufficient conditions. It rather seeks to illuminate central conceptual features of the virtue.

Finally, let me turn to the "facts." My preference is to not presuppose any account of "facts" in (H1), but rather to keep the proposal as ecumenical as possible. That 2 + 2 = 4, that the Earth is round, and that I am not a unicorn are all "facts" on a standard usage of that term. Facts have to be the case and have to capture the way the world really is. But that is compatible with facts being abstract or concrete objects, propositions or the referents of propositions, and the like.

It is not a fact that the Earth is flat. But it is a fact that people used to *believe* that the Earth is flat (and some people still do!). This is a perfectly natural way to talk about facts. But there are limits. My usage does not allow for "alternative facts," to use the expression which originated with Donald Trump's advisor Kellyanne Conway in 2017. There is only one way the facts are.

(H1) is formulated in terms of the facts, not in terms of the agent's *beliefs about* the facts. Which is the right way to go? In other words, does honesty

---

6 For a broadly similar proposal, see Smith 2003: 518, 520. As Smith writes, "Honesty is a refusal to fake reality. It is a person's refusal to pretend that facts are other than they are, whether to himself or others" (518).

An anonymous reviewer asked why I didn't develop the account in terms of "lack of deception." The main reason is that I am trying to provide a more informative account than that would end up being. An additional reason depends upon what one thinks about bald-faced lies, theft, and cheating. If they can still be instances of a failure of honesty even though there is no intention to deceive, then a "lack of deception" approach will be inadequate.

require reliable epistemic access to the facts, or is it enough to just not distort the facts *as the agent takes them to be*?

The latter option seems more plausible to me. Here are two cases that I find persuasive in illustrating why:

> *The Flat Earth Society*. As a member of the Flat Earth Society, Samantha sincerely believes that the Earth is flat. One day she is asked by a friend about the shape of the Earth, and to keep her own beliefs a secret, Samantha tries to mislead her friend and replies that the Earth is round. She succeeds and her friend now assumes that Samantha believes the Earth is round.
>
> Now suppose instead that Samantha is forthright. She tells her friend that she believes the Earth is flat, and has no intention to mislead her friend at all.

It seems to me that in the first version Samantha succeeded in lying, and exhibited a failure of honesty in this case. In the second version, it seems to me that, even though her belief is false, Samantha exhibits honesty in this case.

> *Not in the Library*. Saul tells his mother that he was studying at the library last night, with the intention of misleading her about what he was really doing. He believes that he had actually spent the night at Rachel's apartment. Unbeknownst to him, the person he was spending time with was not named "Rachel" and it was not her apartment.
>
> Suppose instead that, with no intention of misleading her, Saul told his mother that he was spending time at Rachel's apartment last night.

It seems to me that in the first version Saul succeeded in lying, and exhibited a failure of honesty in this case. In the second version, it seems to me that, even though his belief is false, Saul exhibits honesty in this case.[7]

The lesson I take away from cases like these is that both honesty and dishonesty are not tied down to veridical representations of the facts.[8] What we have are different possibilities being exhibited:

(False Belief + False Assertion) leading to Honest Action

---

7 Note that this discussion is about isolated actions, not about the person's trait of honesty or dishonesty. So even if Samantha and Saul failed to exhibit honesty in the first version of the cases, that is entirely compatible with their still being honest people in general. Nevertheless, (H1) has a story to tell about what it is to exhibit or fail to exhibit honesty in a given instance of behavior, and if that story is problematic (as I am suggesting it is), that is grounds for revising (H1) itself. Thanks to an anonymous reviewer for suggesting I clarify this point.

8 For related discussion, see Fried 1978: 58.

(False Belief + True Assertion)	leading to	Dishonest Action
(False Belief + True Assertion)	leading to	Dishonest Action

It is already obvious that there are cases in which we find these combinations:

(True Belief + True Assertion)	leading to	Honest Action
(True Belief + False Assertion)	leading to	Dishonest Action

So given that honest actions don't require that the beliefs of the agent in question be true, we need to revise (H1). In other words, we need an account which ties honesty to subjective representations of the facts, not to the objective facts themselves:

(H2) The virtue of honesty is, centrally, a character trait concerned with reliably not intentionally distorting the facts as the agent sees them.[9]

Even though it will undergo additional refinement elsewhere,[10] (H2) is the core account of the virtue of honesty that I wish to defend.

As a *character trait*, honesty is a set of psychological dispositions which, when activated, give rise to thoughts and feelings that, in turn and other things being equal, lead the person in question to reliably not intentionally distort the facts as she sees them. What this looks like more specifically will depend on what kind of behavior we are talking about. To illustrate, let's take lying first. When (H2) is applied to lying, we get:

> *Lying*: An honest person reliably does not intentionally distort the facts as she sees them by telling lies to others. Nor would she distort the facts about herself in lying to herself either.[11]

If Smith tells his teacher that the dog ate his homework, when Smith in fact never bothered to do it in the first place, then he is intentionally distorting the facts about his homework with the intention of trying to deceive his teacher. Indeed, if the lie is successful, Smith will have distorted the facts in more than one way. His teacher will now believe that the dog ate his homework. And the teacher will now believe *that Smith believes* that the dog

---

9 There are the facts as the person consciously believes them to be, and there are the facts as the agent really believes them to be, but does not consciously recognize due to self-deception or the like. (H2) needs to be refined to take into account such a distinction. See Miller in progress for more details.

10 Miller in progress.

11 For an approach along these lines, see Bok 1978: chapter two. See also MacIntyre: "Truthful persons...do not misrepresent themselves to others as liars do, with regard to the relationship of their beliefs and their intentions to their assertions" (1994: 314).

ate his homework. Both of these beliefs, however, are false.[12] We see similar distortions at work when a media outlet lies about a political candidate's past behavior, or when a political candidate himself lies about that behavior. They are intentionally distorting the facts. Cases of lying fit comfortably with the framework outlined here.

The homework example serves to bring out just how distorting successful lies can be. Obviously with respect to what is being stated, they aim to distort the facts by the lights of the person who is lying. But in the process they also, if successful, distort the facts with respect to how the audience views the beliefs of the liar. Indeed, when the liar says something that, unbeknownst to him, is actually *true*, then there is no distortion of the facts in the audience's mind with respect to the content of what has been said. But there still remains the distortion that comes from the audience forming mistaken beliefs about what the speaker really believes. This arises from the liar intentionally distorting or misrepresenting his own psychology.

In addition to lying, the same goes for misleading others:

> *Misleading*: An honest person reliably does not intentionally distort the facts by her own lights by withholding important information, telling half-truths involving misleading details, or acting in such a way intentionally so as to get others to arrive at a false belief.

Consider this case:

> *The Cheating Spouse*. It is Sunday morning, and a wife asks her husband, "Where were you last night?" Her husband replies, "I was out with the guys at Freddies Bar." He was indeed at the bar from 10–11pm. Afterwards, though, he went back to the apartment of someone he met at the bar.

Now strictly speaking there is no distorting of the facts in this reply. He *was* out with the guys at Freddies Bar. But the distorting comes with the inference that the speaker intentionally wants his spouse to draw. His hope is that she will conclude, "He was *only* out with the guys at Freddies Bar." That clearly distorts the facts.

This case serves to illustrate that the distortion need not be limited to the literal content of what is said, but can include the manner and context

12 As MacIntyre writes, "successful liars necessarily deceive us not only about the subject matter about which they lie, but also about their own beliefs and about their intention in asserting what they assert falsely, and indeed about their further intention to conceal this intention from us" (1994: 313–314). See also Tollefsen 2014: 20, 47.

In our example, the teacher may also believe that Smith intends for the teacher to believe that the dog ate his homework. This, though, is a true belief. For relevant discussion, see Guenin 2005: 181.

in which it is said too. If some of his co-workers ask him, "Which bar did you go to last night?" and he replies, "I was out with the guys at Freddies Bar," there is nothing distorting here with respect to either the content or the manner of delivery. But if he uttered the exact same words in a different context with his wife while also withholding some important information, and he thereby aimed at giving a true response that he hoped would lead her to arrive at a false conclusion, then there is something clearly distorting going on.

Also, misleading others need not be limited to verbal behavior. Painting over the rust on a used car before trying to sell it, also counts as misleading.[13] And clearly it involves intentionally distorting the facts.

The approach can handle some non-standard cases too, such as intentionally making a false statement in order to get someone else to believe something true:

> *The Skeptical Friend*. A's friend B is very skeptical of what A has to say about important matters. A knows about this skepticism. So one day he tells B that "Pluto is still considered one of the nine planets by astronomers." A knows this is false, but hopes to get B to believe the opposite, which is true. Low and behold, B does form the belief that Pluto is now no longer considered to be one of the nine planets by astronomers.

It is not clear whether this counts as a lie or a case of misleading or some third category.[14] But it seems clear that it is a failure of honesty. That is captured by A's intentionally distorting the facts.

One final note. The "intentionally" is important in cases involving misleading others. Suppose that the same person had said in a clear voice, "I was out with the guys at Freddies Bar," but his wife misheard him as saying "I was out with the guys at Froggies Bar." Then there is a distortion of the facts involved, to be sure. But it isn't an intentional distortion on the husband's part, and so with respect to the name of the bar, it does not count as a case of dishonesty.[15]

Lying and misleading fit comfortably with the account in (H2). Other failures of honesty, such as cases of cheating, stealing, and promise-breaking, introduce interesting complexities, and I have discussed them at length

---

13 Carson 2010: 57.

14 For it not being a lie, see Guenin 2005: 183. For Augustine on cases like these, see Griffiths 2004: 28–29 and Decosimo 2010: 664–665. For related discussion, see Tollefsen 2014: 15–16.

15 For related discussion, see Carson 2010: 47.

elsewhere.[16] In the remaining sections of this article, I take up two important challenges to the account.

2. Nazi at the Door Cases

If (H2) or one of its close variants is meant to offer necessary conditions for honesty, then it looks like any distortion of the facts is going to count as a failure to be honest, at least in that one case. But perhaps there are cases where it is compatible with the virtue of honesty *to* intentionally distort the facts. Lying provides the most straightforward and widely discussed cases. Despite earlier arguments by Augustine and Kant, as well as some recent work,[17] most contemporary philosophers seem to hold that situations arise in life in which lying is morally permissible and even morally obligatory. Lying to the Nazis in order to protect the Jews you are hiding in your house is the standard example. Kant's example of lying to the ax-murderer to save his would-be victim is another.[18]

The same possibility of morally permissible cheating arises as well, such as the following:

> *The High-Stakes Game*. The well-being of a child hangs on the outcome of a card game. If Chase wins, he will be able to buy the child from the sex-traffickers and bring her back to her family. If he loses, she will be taken away and it will be very hard to ever find her again. Chase is an expert card sharp. When he tries to play the game fairly, he starts to lose badly. His only chance of winning is to start cheating, which he does. As a result, he wins the game and buys the child's freedom.

Chase clearly cheats, but arguably his doing so is morally permissible. There were also cases of morally permissible stealing, such as this:

> *The Hurricane*. To stop a hurricane from destroying their house with their children inside, a couple might steal some unused plywood in their neighbor's yard, even though it will be unreturnable after the storm hits.

Even though the couple is fully aware that it was rightfully the neighbor's plywood, it seems that competing considerations could morally justify their taking it in a situation like this one.

---

16 See Miller in progress. There I also consider the relationship between bullshit and honesty.

17 See Finnis 1980, Murphy 1996, Garcia 1998, and Griffiths 2004.

18 For additional examples, see LaFollette and Graham 1986: 8–13, Guenin 2005: 207, and Stokke forthcoming.

Let's assume in this section that there are cases of morally permissible lying, cheating, stealing, and so forth. Can my approach accommodate them? A natural revision to make to (H2) is something like this:

(H3) The virtue of honesty is, centrally, a character trait concerned with reliably not intentionally distorting the facts as the agent sees them, so long as the agent does not take it to be morally permissible to do so.

Note that, in the spirit of (H2), this is a subjective way of developing the exception. On a more objective approach, we could say:

(H3*) The virtue of honesty is, centrally, a character trait concerned with reliably not intentionally distorting the facts as the agent sees them, so long as it is not morally permissible to do so.

It would then be the job of different ethical theories to tell us when it is or is not morally appropriate to lie, cheat, or steal. Standard utilitarianism will have a very different answer to give than standard forms of virtue ethics or divine command theory.[19]

Matters are more complicated than this, however. Let me focus just on the case of lying to focus the discussion, but what I say in the remainder of this section generalizes. We should distinguish between two different views about how honesty and morally permissible lying are related:

(i) The virtue of honesty does not apply to certain cases of lying, say lying to the Nazi in order to protect a Jewish family.
(ii) Lying in certain cases, such as to the Nazi in order to protect a Jewish family, is still a failure of honesty, but it is all-things-considered morally permissible.

One way to develop the thought behind (i) is along the following lines. Suppose that in order to count as a lie, it has to be the case that one's intended audience has a right to know the truth. Without that right, there is no lie, *even if* I say something I know to be false with the intention of deceiving my audience.[20] In the case in question, the thought would be that the Nazi has no right to know the whereabouts of the Jews he intends to harm. Hence telling the Nazi a bogus location of where the Jews are, would not fall under the purview of the virtue of honesty in this case since it does not count as a genuine lie.

---

19 For relevant discussion, see Fried 1978: chapter three, Gert 1998: chapter eight, Garcia 1998: 521, and Carson 2010.

20 For relevant discussion, see Bok 1979: 14–15, Carson 2010: 18–20, and Tollefsen 2014: 25–30. Benjamin Constant is said to hold the view above (MacIntyre 1994: 341), as did Grotius (Tollefsen 2014: 6).

Another way to develop the thought behind (i) is to grant that telling the Nazi the wrong location would indeed be a lie, but that other considerations about the well-being of the Jews simply silence or eliminate consideration of the moral status of lying.

Initially when I was revising (H2), I assumed that something like this option in (i) was the case. But that is not obvious. According to (ii), someone hiding the Jews would indeed be telling a lie, the lie is morally justified, *and* the person is being dishonest. It is just that other virtues, such as benevolence or non-malevolence, take greater priority in such instances, and end up justifying lying all-things-considered.[21]

If we go with option (ii), then no revision to (H2) is needed after all.[22] We could say that in the Nazi case there are still normative facts pertaining to the *pro tanto* wrongness of lying, but those normative facts are being outweighed by other normative facts having to do with benevolence, for instance.[23]

Which option is more plausible? I do not have to take a stand, and could leave it up to the reader to decide. If it is (i), then a revision to (H2) has been provided above to accommodate it. If it is (ii), then no revision is needed.

For what it is worth, let me report that my sympathies have come to rest with (ii) and the claim that even morally justifiable lying (and cheating, stealing, and the like) is a failure of honesty. For one thing, it seems intuitively obvious to me that the person would be telling a genuine lie in intentionally giving a false location to the Nazi at the door. This intuition is fallible, and

---

21 I put this in terms of virtues, but the point can be put more neutrally just in terms of other morally relevant considerations which take greater priority than not intentionally distorting the facts. One implication of putting the point in terms of virtues, is that it seems like it would lead to a denial of the unity of the virtues thesis. Like most philosophers working on virtue, I find this to be an implication that is perfectly acceptable. Thanks to an anonymous reviewer for pointing it out.

22 Thanks to David Carr for relevant discussion.

23 Following Rosalind Hursthouse, one way to develop this thought further is in terms of what she calls resolvable moral dilemmas, which despite being resolvable often are such that "the overridden requirement retains its force in some way, so regret, or perhaps the recognition of a new requirement, are still appropriate" (1999: 44). Similarly, Christopher Tollefsen describes "the sense in which our brokenness and sinfulness – indeed, not just our own, but that of the world – makes it impossible for us to avoid sin; there are genuine cases of necessity in which one must act in a way that is imperfect, guilty sinful – yet nevertheless, to repeat, one *must* act in that way (Tollefsen 2014: 62, emphasis his; see also 61, 68, 71–72).

Note, though, that the Nazi case is not a moral dilemma in the strict sense of being required to perform two actions which cannot be jointly performed and where the moral requirements do not outweigh each other. Furthermore, it does not rise to the level of what Hursthouse calls a "tragic" moral dilemma, which can be resolvable or irresolvable on her view. In those dilemmas, "a virtuous agent cannot emerge with her life unmarred" (79).

there may be powerful theoretical considerations that could force me to give it up, but I have yet to see them.[24]

Leaving aside the no-right-to-truth option, more generally it seems that (ii) is more intuitively plausible than (i). If I am asked – is lying to the Nazi a case where honesty just did not apply at all, or is it a case where the person did something dishonest but it was still justified overall? – I find the second option more compelling. In a similar vein, Tom Carson gives us this medical case:

> Suppose that a man has just had open heart surgery and is temporarily in a precarious state of health. His surgeon says that he must be shielded from any emotional distress for the next few days. Unbeknownst to the patient, his only child, Bob, has been killed in an automobile accident. When the patient awakens after the surgery, he is surprised that Bob is not there and asks, "Where is Bob?" You fear that in his condition, the shock of learning about Bob's death might cause the man to die. So you lie and say that his son has been delayed...This seems to be a case of morally permissible lying that violates someone's right to know the truth.[25]

And I might add, it seems to be a case where you would be doing something dishonest. While I do not have any data to support this, I suspect my intuitions are in line with ordinary discourse and folk psychology here.

A third reason is theory-driven – if failing to be honest is, at its core, a matter of intentionally distorting the facts, then lying to the Nazi or the ax-murderer is no less a matter of intentionally distorting the facts than is morally prohibited lying. There is the same basic failure when it comes to

---

24 As Tom Carson writes about this case, "Ordinary language counts the example in question as a case of lying. There is a strong presumption against any definition of lying so much at odds with ordinary language. Using the term 'lying' in accordance with this definition is likely to engender confusion" (2010: 19). He provides additional arguments against this approach as well at 2010: 19–20, and see also Tollefsen 2014: 25–30, 90–92, and chapter three.

25 Carson 2010: 19–20. Similarly, cases of lying under duress support (ii) as well. Here is one such case from Stuart Green:

> Imagine that A, while having a gun held to his head by B, is forced to lie to C, who is on the other end of the telephone. A has done something wrongful; he has misled C, and he has done so intentionally; he has acted unjustifiably. But A has acted under duress. Although A's act itself was wrongful, most of us would agree that he should not be blamed for it – that A's conduct, in other words, should be excused (2006: 84).

I would only add to its being wrongful that A's action was also dishonest, yet still all-things-considered morally permissible. Note that here too this does not require a revision to H2, since the duress bears on the blameworthiness and the all-things-considered moral permissibility of the lie, but not on its being a failure of honesty. Thanks to an anonymous reviewer for encouraging me to clarify this.

conveying what reality (as the person sees it) is like. So it is hard to see why the scope of honesty would fall short of these lies.

Finally, a fourth reason is that (ii) parallels how other virtues work. Suppose just for the sake of discussion that torturing a terrorist in order to disclose the location of a bomb is all-things-considered morally justified. Nevertheless, the act itself is cruel, and is a failure of non-malevolence. The virtue term applies, but the relevant considerations – we are supposing – get outweighed.

But I don't have to close off options here. While I personally prefer (ii), one of my main goals is to get a number of possible views out on the table.

3. The Challenge of Radically Mistaken Beliefs about Moral Norms

A different challenge to (H2) has to do specifically with its treatment of cheating and stealing. Consider cases such as this:

> *The Fight*. Atticus has been forcibly enslaved and thrown into the coliseum to fight against the Roman gladiator for the entertainment of the crowd. Atticus is not given a fair chance; he only has a wooden shield to use against the armor and sword of the gladiator. But he manages to sneak in a small piece of metal which, at a key moment in the battle, he uses to cut the gladiator's throat. This is against the rules, and the crowd boos and calls Atticus a cheater.

Did Atticus cheat? In a sense, he did. He intentionally violated the rules governing this activity. This is the "factual" or "descriptive" sense in which Atticus cheated, even though objectively speaking he might not have done anything wrong. Hence it might seem on initial inspection that according to (H2), Atticus's killing the gladiator constitutes a failure of honesty since he intentionally distorted the facts of what constitutes participating in this activity.

But that might seem implausible. The rules set up for this fight were blatantly unjust. Why should Atticus's honesty be faulted when he fails to follow those rules, thereby saving his own life in the process?[26]

Fortunately (H2) doesn't have to have this implication. For Atticus was not distorting the facts with respect to what he considered to be a fair fight. He was distorting the facts according to the people who controlled him, but not according to his own lights.

Cases of stealing present similar challenges for (H2). Take by way of illustration an abolitionist helping to secretly rescue a slave from a plantation.

---

26 As Green writes about another, related case: "it would not have been cheating for a girl in Afghanistan under the Taliban to violate the law that made it a crime for her to attend school, since the law itself was surely unjust and issued by an illegitimate authority" (2006: 63).

Again there is a "descriptive" sense of stealing at work here. Given the prevailing social norms and laws of the time and place, what the abolitionist was doing would count as theft. Yet it is also hard to call such an abolitionist "dishonest" even though he was distorting the "facts" in society at the time about property rights.

However, the abolitionist presumably did not accept that what were said to be the "facts" about property ownership and slavery at the time, *really were* facts. So in helping to free the slave, he was not distorting the facts by *his* lights. On (H2), this would not count as failing to exhibit honesty.

But when we turn to the famous case of Huck Finn in Mark Twain's novel, matters are more complicated. During the time of slavery in the American South, Huck is faced with a choice between turning his friend Jim in to the authorities as a runaway slave, or helping Jim to escape. He ends up doing the latter. But Huck clearly thinks that what he is doing is morally wrong, that it constitutes stealing from Jim's "master." He judges that he should turn Huck in, and when he doesn't do so, he considers himself a bad boy. Yet clearly he is not. Twain depicts Huck Finn as more practically wise and perceptive – in a word, more virtuous – that many of the adults of that society.[27]

Does (H2) give the wrong verdict here? Since it ties honesty to not distorting the facts *as the person sees them*, and since Huck judges the facts (both descriptive and normative) to require him to turn Huck in, (H2) seems to imply that Huck failed to be honest in this one instance. Even if that evaluation is compatible with Huck Finn's also being highly compassionate, caring, and even honest in many other situations, it still might be hard to accept.[28]

Fortunately, though, (H2) does not have to imply this about Huck. For while Huck was distorting certain "facts" by his lights, he was not distorting other ones. He was responding to his experience of what his friendship with Jim means to him, and how he had come to see Jim as a genuine person about whom he cared a great deal. Those strike me as facts too, even if they were not part of his consciously formed moral judgment.

So in helping Jim, Huck was distorting certain facts by his own light, but not distorting other facts. And the ones that were truly more important to him were the ones that went against his conscious moral judgment to turn Jim in. So relative to the facts about which he cared the most, he was not failing to be honest.

This can lead to another revision to (H2), where someone's honesty is relativized to various kinds of norms:

---

27 Here I have been helped by Arpaly 2002.

28 As noted in footnote 7, even though (H2) is an account of the trait of honesty, it still has implications for what it is to succeed or fail at performing a particular honest action. And if it would suggest that Huck Finn failed to exhibit honesty in this one case, then many will see that as a drawback of the account.

(H4) The virtue of honesty is, centrally, a character trait concerned with reliably not intentionally distorting the facts as the agent sees them, including what the agent takes (either consciously or unconsciously) to be the normative facts thought to be relevant in a given situation. If the agent takes there to also be opposing normative facts which bear on the situation, then she may be honest or fail to be honest in relation to each set of normative facts. In addition, she may be honest or fail to be honest in relation to the normative facts all-things-considered.[29]

It seems intelligible to say that relative to the societal norms of his day which Huck used in forming his conscious judgment, he failed to exhibit honesty by subsequently helping Jim to escape. But relative to his more deeply held norms of friendship and caring for Jim, his behavior did not fail.

Similarly, returning to the case of Atticus and the gladiator, (H6) can capture what we called the "descriptive" sense in which Atticus was a cheater. Relative to the social norms for fights like these in ancient Rome, his behavior distorted the normative facts. He was not allowed to bring a weapon into the arena. But (H6) can also capture the sense in which he did not do anything dishonest, namely relative to what he took to be the normative facts having to do with justice and with his own self-preservation.

(H4) also serves to emphasize that even though honesty is still being understood using a subjective approach to thinking about the facts, for the agent in question they need not be part of her conscious awareness in the moment.

As far as this paper in concerned, (H4) is the final revision of the account of honesty that will be considered.

4. Conclusion

In this paper I have begun the work of developing and defending an original account of the virtue of honesty. Even though the account went through multiple revisions, there is still much more work that needs to be done. Motivation, for instance, is not addressed by (H4) at all. But hopefully important progress has been made in these largely uncharted philosophical waters.[30]

---

29 Connecting failures of honesty to distortions of the normative as well as the descriptive facts, has a number of important and controversial implications that I have explored in some detail in Miller in progress. One such implication, for instance, is that every case of morally wrong behavior which is taken by the agent to be morally wrong, will also count as dishonest.

30 I am very grateful to Voin Milevski for inviting me to be a part of this special issue and to two anonymous reviewers for very helpful comments. Work on this paper was supported by a grant from the John Templeton Foundation and from the Templeton Religion Trust. The opinions expressed here are those of the author and do not necessarily reflect the views of these Templeton Foundations.

Works Cited

Arpaly, Nomy. (2002). *Unprincipled Virtue*. New York: Oxford University Press.

Bok, Sissela. (1978). *Lying: Moral Choice in Public and Private Life*. Pantheon Books.

Bratman, Michael. (1987). *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Caouette, Justin and Carolyn Price. (2018). *The Moral Psychology of Compassion*. Lanham: Rowman and Littlefield.

Carson, Thomas. (2010). *Lying and Deception: Theory and Practice*. Oxford: Oxford University Press.

Decosimo, David. (2010). "Finding Augustine's Ethics of Public Lying in His Treatments of Lying and Killing." *Journal of Religious Ethics* 38: 661–697.

Driver, Julia. (1989). "The Virtues of Ignorance." *Journal of Philosophy* 86: 373–384.

Finnis, John. (1980). *Natural Law and Natural Rights*. Oxford: Oxford University Press.

Fried, Charles. (1978). *Right and Wrong*. Cambridge: Harvard University Press.

Garcia, J. L. A. (1998). "Lies and the Vices of Deception." *Faith and Philosophy* 15: 514–537.

Gert, Bernard. (1998). *Morality: Its Nature and Justification*. Oxford: Oxford University Press.

Guenin, Louis. (2005). "Intellectual Honesty." *Synthese* 145: 177–232.

Green, Stuart. (2006). *Lying, Cheating, and Stealing: A Moral Theory of White-Collar Crime*. Oxford: Oxford University Press.

Griffiths, Paul. (2004). *Lying: An Augustinian Theology of Duplicity*. Grand Rapids: Brazos Press.

Hursthouse, Rosalind. (1999). *On Virtue Ethics*. Oxford: Oxford University Press.

LaFollette, Hugh and George Graham. (1986). "Honesty and Intimacy." *Journal of Social and Personal Relationships* 3: 3–18.

MacIntyre, Alasdair. (1981). *After Virtue*. Notre Dame: University of Notre Dame Press.

MacIntyre, Alasdair_. (1994). "Truthfulness, Lies, and Moral Philosophers: What Can We Learn from Mill and Kant?" *The Tanner Lectures on Human Values* 16: 308–361.

Mele, Alfred. (1995). *Autonomous Agents: From Self Control to Autonomy*. Oxford: Oxford University Press.

Miller, Christian. (2017). "Honesty," in *Moral Psychology, Volume V: Virtue and Character*. Ed. Walter Sinnott-Armstrong and Christian B. Miller. Cambridge: MIT Press, 237–273.

Miller, Christian. (in progress). *Honesty: The Philosophy and Psychology of a Neglected Virtue*.

Miller, Christian and Ryan West. (Forthcoming). *Integrity, Honesty, and Truth Seeking*. Ed. Christian B. Miller and Ryan West. New York: Oxford University Press, forthcoming.

Murphy, Mark. (1996). "Natural Law and the Moral Absolute Against Lying." *American Journal of Jurisprudence* 41: 81–101.

Smith, Tara. (2003). "The Metaphysical Case for Honesty." *The Journal of Value Inquiry* 37: 517–531.

Stokke, Andreas. (Forthcoming). "Lies, Harm, and Practical Interests." *Philosophy and Phenomenological Research* https://doi.org/10.1111/phpr.12439.

Tollefsen, Christopher. (2014). *Lying and Christian Ethics*. Cambridge: Cambridge University Press.

Wilson, Alan. (2018). "Honesty as a Virtue." *Metaphilosophy* 49: 262–280.

Caj Strandberg
Department of Philosophy, Classics,
History of Art and Ideas
University of Oslo, Norway

Original Scientific Paper
UDC 17.022.1:165.022/.023
17.03:161/164

INTERNALISM AND THE FREGE-GEACH PROBLEM*

Abstract. *According to the established understanding of the Frege-Geach problem, it is a challenge exclusively for metaethical expressivism. In this paper, I argue that it is much wider in scope: The problem applies generally to views according to which moral sentences express moral judgments entailing that one is for or against something, irrespective of what mental states the judgments consist in. In particular, it applies to motivational internalism about moral judgments. Most noteworthy, it applies to cognitivist internalism according to which moral judgments consist in motivating beliefs. Hence, in order for a metaethical view to evade the Frege-Geach problem, it should avoid stating that moral judgments are motivating.*

Key words: *moral judgment, motivation, Frege-Geach problem, embedding, internalism, externalism, cognitivism, expressivism, hybrid view, ecumenical view, besire*

1. Introduction

The Frege-Geach problem—henceforth 'the F-G problem'—is without doubt one of the most discussed arguments in metaethics. According the traditional understanding of the problem, it provides a challenge exclusively for expressivism. The fundamental point is thought to be that expressivism is unable to account for the meaning of moral sentences when they occur in embedded contexts, since this view claims that such sentences express non-cognitive states. In this paper, I argue that the F-G problem should not be understood to concern what kind of mental states moral sentences express. Rather, it concerns whether the mental states that moral sentences express entail that one is for or against something, what I will refer to as 'approval' or 'disapproval'. The upshot of this finding is that the F-G problem is much

* The first version of this paper was written already in 2008. After some unsuccessful attempts to have it published, I kept it in the drawer until I received the generous invitation to contribute to the present issue. In one of the journals in which I tried to get the paper published, a suspiciously similar argument later occurred. I am particularly grateful to Gunnar Björnsson, John Eriksson, and Ragnar Francén for comments on early versions of the text.

 DOI: 10.5937/BPA1932067S

wider in scope than normally thought: It applies to views according to which moral sentences express moral judgments that entail approval or disapproval, quite irrespective of whether they consist in non-cognitive states or not. In particular, the problem applies to motivational internalism about moral judgments, which is the most well-known instance of this kind of view.

In the next section, I explain why the F-G problem constitutes a challenge for expressivism. In Section **3**, I argue that is plausible to think that the problem is wider in scope in the way indicated above. In Section **4**, I make a distinction between two kinds of internalism: *state internalism* and *object internalism* depending on whether it is the mental state or the object of the state that explains motivation. In Section **5**, I discuss *state internalism*. There are three main types of state internalism: *non-cognitivist internalism*, according to which a moral judgment consists merely in a non-cognitive state; *hybrid internalism*, according to which it consists in both a non-cognitive and a cognitive state, and *sui generis internalism*, according to which it consists in a *sui generis* motivating and representational state. It is argued that all three views are subject to the F-G problem. In Section **6**, I discuss *object internalism* in the form of *cognitivist internalism*. According to this view, a moral judgment consists in a cognitive state understood as a motivating belief. It is argued that this view also is susceptible to the F-G problem, in spite of stating that moral judgments consist merely in beliefs. In Section **7**, I explain that a certain weak version of internalism is not subject to the F-G problem. Finally, in Section **8** I draw three metaethical lessons from the previous discussion.

2. Expressivism and the Frege-Geach Problem

Let us start with adopting some familiar terminology that will enable us to formulate various metaethical claims which will be discussed in what follows.

Think of a well-formed English sentence. The sentence has a certain conventional meaning that constitutes its *semantic content*. Assume that a person asserts or accepts the sentence. It is then plausible to assume that she is in a certain mental state that corresponds to the content of the sentence. We might say that the sentence, by virtue of its meaning, *expresses* the mental state in question. More precisely, what a sentence expresses can be understood as the mental state that a person needs to be in, in order for it to be compatible with the meaning of the sentence that she accepts or asserts it.[1] As regards ordinary fact stating sentences, this is straightforward: The semantic content of the sentence 'It is raining' is the proposition: it is raining. The sentence expresses the belief that it is raining, i.e. a belief which has the mentioned proposition as its object.

1 See e.g. Schroeder (2008): Ch. 2. Cf. Ridge (2003): 563–574, and Kalderon (2005): Ch. 2.

Think now of a moral sentence such as 'It is wrong to ϕ'. The sentence has a conventional meaning that constitutes the semantic content of the sentence. In case a person understands the meaning of the sentence and accepts or asserts it, she finds herself in a certain mental state corresponding to the content of the sentence. We might adopt a common metaethical convention and refer to this state as a *moral judgment*. Accordingly, the sentence expresses, by virtue of its meaning, a moral judgment.[2] Metaethical views can now be formulated both in terms of the contents of moral sentences and in terms of the mental states that these sentences express.

Expressivism is a claim about the meaning of moral sentences. Understood as a thesis about what moral sentences express, it can be formulated thus:

> *Expressivism*: A moral sentence, such as 'It is morally wrong to ϕ', expresses, by virtue of its meaning, a moral judgment that consists in a non-cognitive state in relation to ϕing.

Expressivism can also be formulated in terms of the semantic content of moral sentences. According to expressivism, moral sentences do not, in contrast to ordinary fact stating sentences, express beliefs. There are consequently no moral beliefs that have moral propositions as their objects where these propositions constitute the contents of moral sentences. Rather, on this view the contents of moral sentences consist in the non-cognitive states they express. In Mark Eli Kalderon's words, on expressivism 'the content of a moral sentence wholly consists in non-cognitive attitudes conveyed by its utterance' and this view thus reduces the contents of moral sentences to what they express.[3] Thus formulated, expressivism amounts to the following: A moral sentence like 'It is wrong to ϕ' has a semantic content that consists in a non-cognitive state in relation to ϕing, i.e. the non-cognitive state which the sentence expresses.

Expressivism claims that moral judgments consist in a particular type of mental states: non-cognitive states. There are presumably a number of different types of non-cognitive states, such as desires, emotions, and wishes. Moreover, there are different metaphysical theories about how this type of mental states should be characterized. However, it is generally agreed that they have two features. *First*, a non-cognitive state does not represent as certain state of affairs as being the case. It thereby contrasts with a cognitive state, primarily beliefs, which has this function. Second, a non-cognitive state is such that if a person is in this type of state, she is *for* or *against* something. In what follows, I will formulate this aspect by saying that she *approves* or *disapproves* of something. Thus, non-cognitive states have an essential

---

2 I take 'moral judgment' to be neutral between cognitivism and non-cognitivism. On the former view, it consists in a cognitive state (like a belief); on the latter, it consists in a non-cognitive state (like a desire).

3 Kalderon (2005): 53. Cf. Blome-Tillman (2009): 279–285.

feature: They entail approval or disapproval. However, this is compatible with the possibility that they share this feature with other mental states. This will be important later on.

Let us now consider the F-G problem. It is helpful to describe it in three steps where the second step is the crucial one.[4] First, consider a freestanding sentence: (**1**) 'It is wrong to lie'. According to expressivism, (**1**) expresses a non-cognitive state such that a person who finds herself in this state disapproves of lying: she is against lying. Expressivism gets support from the fact that it seems very plausible that a person who accepts (**1**) disapproves of such actions. Second, consider a complex sentence where (**1**) occurs embedded: (**2**) 'If it is wrong to lie, it is wrong to get one's little brother to lie'. It seems evident that a person might accept (**2**) without disapproving of lying, since she need not think that lying is wrong. Third, a sentence has the same meaning irrespective of whether it occurs freestanding or embedded.[5]

What I consider as the basic point in the F-G problem amounts to the following when applied to expressivism. According to expressivism, a freestanding sentence such as (**1**) expresses, by virtue of its meaning, a non-cognitive state, which entails that a person who accepts (**1**) disapproves of lying. However, it seems that a person who accepts a complex sentence, such as (**2**), in which (**1**) is embedded, need not disapprove of such actions. Hence, it appears that a person who accepts (**2**) need not be in the mentioned non-cognitive state. Expressivists then owe us an explanation as to how a moral sentence, such as (**1**), can have the same meaning when it occurs freestanding and when it occurs embedded, such as in (**2**).[6]

In contemporary metaethics, it is commonly stressed that expressivists have the problem of explaining how the meaning of complex sentences can be a function of the meaning of the sentences that constitute their parts. The most common illustration concerns logically valid arguments. Consider:

(**1**) It is wrong to lie.

(**2**) If it is wrong to lie, then it is wrong to get one's little brother to lie.

(**3**) Therefore, it is wrong to lie.

Clearly, (**3**) logically follows from (**1**) and (**2**). However, in order for (**3**) to follow from (**1**) and (**2**), it is appears that the antecedent in (**2**) needs to have the same meaning as (**1**). Thus, expressivists owe us an explanation as to how such arguments can be valid. More generally, they need to explain how complex sentences, such as (**2**), get their meaning from their parts, such as (**1**) and (**3**).

4 Cf. Schroeder (2010): 44–47.

5 Geach (1965): 449.

6 For two early formulations of this problem, see Geach (1960): 221–225, and Searle (1962): 423–432. For some recent and clear accounts, see e.g. Sinnott-Armstrong (2000): 677–693; Kalderon (2005): 52–66, and Schroeder (2010), Ch. 3, 6, and 7.

3. Generalizing the Frege-Geach Problem

In what follows, I would like to draw attention to an aspect of the F-G problem that seems to have gone unnoticed in the debate.[7] The defining characteristic of expressivism is that moral sentences express a certain type of mental states: non-cognitive states. However, the F-G problem does not refer to the claim that moral sentences express a *particular type of mental states*. Rather, it refers to the claim that moral sentences express mental states *that have a certain feature*: they entail *approval* or *disapproval*. More precisely, it appeals to the fact that a person who finds herself in a non-cognitive state with regard to an action entails that she approves or disapproves of the action in question, that she is for or against it. Thus, it is not the claim that a moral sentence such as 'It is wrong to ϕ' expresses a non-cognitive state which is the root of the problem for expressivism, but rather the claim that the sentence expresses a mental state which has a certain feature: it entails disapproval of ϕing. In other words, it is the 'being for or against' feature that is the real target of the F-G problem, rather than moral judgments consisting in a particular type of mental states.

The fact that expressivism claims that a moral sentence such 'It is wrong to ϕ' expresses a non-cognitive state is *relevant* as to why this view is susceptible to the F-G problem. However, this fact is merely *indirectly* relevant. It is relevant because the fact that a person finds herself in the non-cognitive state in question *entails* that she disapproves of ϕing. It is not directly relevant because the problem does not refer to the non-cognitive state as such, but to a certain feature that is had by such mental states.

To see this clearer, recall the second and crucial step in the F-G problem. Its fundamental point is that a person who accepts (**2**) need not disapprove of lying, not that she need not find herself in a non-cognitive state as regards lying. The fact that a person who accepts (**2**) need not disapprove of lying entails that she need not be in the non-cognitive state that (**1**) is assumed to express. However, this is merely a consequence of the fact that a person finding herself in this non-cognitive state entails that she disapproves of lying. The point does not appeal as such to the expressivist claim that a moral sentence expresses a non-cognitive state.

Importantly, this suggests that the F-G problem might apply to other metaethical views than expressivism.[8] As we have seen, it is not the fact that expressivism claims that a moral sentence such as 'It is wrong to ϕ' expresses a non-cognitive state which is the root of the problem for this view, but rather that the non-cognitive state entails a particular feature: disapproval of ϕing.

---

7 I develop this part of the argument in more detail in Strandberg (2015a): 1–15

8 In Strandberg (2015a): 1–15, I provide a fuller explanation of why metaethicists have been led to think that the F-G problem applies exclusively to expressivism. For another manner in which the problem might generalize, see Eklund (2009): 705–712.

However, as already mentioned, there might be other mental states that entail that one is for or against something. This suggests that other metaethical views, according to which moral sentences express mental states that have this feature, also are subject to the F-G problem.

4. State Internalism and Object Internalism

In the last section, it was hypothesized that the F-G problem can be generalized to metaethical views according to which moral sentences express mental states entailing approval or disapproval. These views have the following claim in common:

> *The Intrinsic Claim*: It is conceptually necessary that, if a person judges that ϕing is morally wrong, then she disapproves of ϕing.

In the remainder of the paper, I will not be concerned with this abstract claim, but with a view that is in the focus of much of the metaethical debate: motivational internalism. There are presumably a number of different types of approval and disapproval, since there are different ways of being for or against something. However, one important characteristic of being for or against something is that one is *motivated* in different manners. As a consequence, it can be hypothesized that the F-G problem is generalizable to internalism according to which moral sentences express moral judgments that involve motivation.

A generic version of internalism can be formulated as follows:

> *Motivational Internalism*: It is conceptually necessary that, if a person judges that it is morally wrong to ϕ, then she is, at least to some extent, motivated to see to it that ϕing is not performed.[9]

Internalism can be formulated as a claim about what moral sentences express: The moral sentence 'It is morally wrong to ϕ' expresses a moral judgment which is such that, if a person finds herself in this mental state, then she is motivated to see to it that ϕing is not performed.

In what follows, I will be concerned with a broader version of internalism than what normally is considered. According to this view, there is a

---

9 For an overview of different types of internalism, see Björnsson et al. (2015): 1–20. For helpful clarifications of particular aspects of internalism and alternative manners of understanding it, see e.g. Cuneo (1999): 361–363; Svavarsdóttir (1999): 163–165; Lillehammer (2002): 1–25; Lippert-Rasmussen (2002): 8–15; Schroeter (2005): 1–23; Tresan (2006): 143–148; Tresan (2009): 51–72; Zangwill (2007): 93–97; C.B. Miller (2008): 233–255; Francén (2010): 117–148; van Roojen (2010): 495–525; Strandberg (2011): 341–369, and Strandberg (2012): 81–91. The literature also includes considerations about the empirical support of internalism. See e.g. Roskies (2003): 52–53; Cholbi (2006): 607–616; Kauppinen (2008): 1–24; Strandberg and Björklund (2013): 319–335, and Milevski (2015): 113–126.

conceptually necessary connection between a person's moral judgment about an action and her general motivation in relation to it, not merely between her moral judgment about her own prospective action and her motivation to perform or not to perform it. Thus, the phrase 'see to it that ϕing is not performed' should be understood to include all types of cases where a person is motivated to hinder ϕing in various manners, e.g. being motivated not to ϕ herself, motivated to hinder others from ϕing, motivated to advice other people not to ϕ, etc.

However, in a fundamental respect I will adhere to the traditional understanding of internalism, since I will be concerned with a view according to which a person's moral judgment is part of what explains her motivation. Thus, a person's moral judgment that it is wrong to ϕ is part of the explanation of why she is motivated to see to it that ϕing is not performed.[10] Thus, I will be concerned with versions of internalism according to which motivation is 'internal' or 'intrinsic' to moral judgments.

We might further distinguish between two versions of internalism of this kind. According to *unconditional* versions of internalism, the necessary connection between moral judgments and motivation holds for every person. According to *conditional* versions of internalism, this connection holds only for those who satisfy a certain condition. In what follows, I will formulate my arguments in terms of the first version in order to avoid unnecessary complications. In Section **7**, I return to this distinction and explain that there are certain forms of unconditional internalism which are not susceptible to the F-G problem.[11]

Internalism, as formulated so far, does not say anything about what it is about a moral judgment which explains that it is motivating. There are basically two alternatives: It might be something about *the kind of mental state* that constitutes a moral judgment, *or* it might be something about the *proposition* that is the object of the moral judgment. Thus, there is a distinction between two types of internalism that will be useful in the ensuing discussion:

> *State Internalism*: (**i**) Motivational Internalism. (**ii**) It is the fact that a person's moral judgment to the effect that it is morally wrong to ϕ involves a kind of mental state that is motivating which explains that she is motivated to see to it that ϕing is not performed.

---

10 According to an alternative version of internalism, we *classify* a judgment as a *moral* judgment only if it is accompanied by motivation, but the moral judgment is not involved in the explanation of the motivation. See Tresan (2006): 143–165, and Tresan (2009): 51–72. Cf. Sneddon (2009): 41–53. My arguments do not affect this version of internalism. I argue against this view in Strandberg (2016): 42–43.

11 According to yet another version of internalism, the necessary connection between moral judgments and motivation does not hold on an individual level, but at a communal level. See e.g. Gert and Mele (2005): 275–283, and Bedke (2009): 189–209. My arguments do not affect this version of internalism.

> *Object Internalism*: (**i**) Motivational Internalism. (**ii**) It is the fact that a person's moral judgment to the effect that it is morally wrong to ϕ has a certain proposition as its object which explains that she is motivated to see to it that ϕing is not performed.

These views are in principle neutral as regards what kind of mental state a moral judgment consists in. However, they naturally connect with two distinct views in this regard.

According to *state internalism*, it is the fact that a moral judgment involves a *kind* of mental states which is characterized by being motivating that explains motivation. On the most common version of this view, moral judgments partly or wholly consist in a particular type of non-cognitive states that motivate to action: *desires*. It is often maintained that there are cognitive states, in the form of beliefs, that can motivate, but this is not something that characterizes beliefs as a *kind* of mental states, as is suggested by the plausible view that not *all* beliefs motivate. On another version of state internalism, moral judgments consist in *sui generis* mental states ('besires'), which are understood as mental states that are neither beliefs nor desires, but which belong to the kind of mental states that is motivating.

According to *object internalism*, it is the fact that a moral judgment has a certain proposition as its object which explains that it is motivating. Object internalism is naturally combined with the view that moral judgments consist in *beliefs*. In case a moral judgment involves a mental state belonging to a *kind* of mental states that is motivating, there would be no need to refer to the propositional object of the state to explain motivation, which suggests that moral judgments consist in beliefs on this view. Moreover, it must be something about moral beliefs that explains why they, as opposed to other beliefs, are motivating. The explanation seems to be that such a belief has a moral proposition as its object.

5. State Internalism and the Frege-Geach Problem

There are primarily three versions of object internalism: non-cognitivist internalism, hybrid internalism, and *sui generis* internalism.

5.1. *Non-Cognitivist Internalism*

The simplest version of state internalism maintains that moral judgments consist in desires:

> *Non-Cognitivist Internalism* (*NCI*): (**i**) Motivational Internalism. (**ii**) A moral judgment consists in a desire.

According to this view, the sentence 'It is wrong to ϕ' expresses a moral judgment which consists in a desire that ϕing is not performed. In order to

explain that a person has such a desire, we need to assume that she has a desire that actions that have a certain feature F is not performed and that she believes that ϕing has F. However, on the present view this latter desire and belief are not part of the judgment that a moral sentence expresses. The moral sentence only expresses a desire with regard to ϕing.

We can now see that *NCI* is subject to the F-G problem. First, consider a freestanding sentence such as (**1**): 'Lying is wrong'. According to *NCI*, the sentence expresses a judgment which consists in a desire that lying is not performed. It follows that a person who accepts (**1**) is motivated to see to it that lying is not performed. Second, consider a sentence in which (**1**) is embedded, such as (**2**): 'If it is wrong to lie, then it is wrong to get one's little brother to lie'. It seems evident that person might accept this sentence without being motivated to see to it that lying is not performed, since she need not think that lying is wrong. Thus, advocates of *NCI* owe us an explanation as to how (**1**) can have the same meaning when it occurs freestanding, as in (**1**), and embedded, as in (**2**).

It should not come as a surprise that *NCI* is subject to the F-G problem, since it entails expressivism which is the traditional target of the argument. However, what is noteworthy is the reason *why NCI* is susceptible to this problem. The reason is not that it claims that a moral sentence such as (**1**) expresses a moral judgment consisting in a non-cognitive state in the form of a desire. The reason is rather that this view entails that such a sentence expresses a moral judgment consisting in a mental state that is motivating. Thus, the explanation why *NCI* is subject to the F-G problem verifies the suggestion in Section **3** that the F-G problem is wider in scope than usually thought.

5.2. Hybrid Internalism

According to *hybrid internalism*, a moral judgment consists in a complex mental state constituted by a non-cognitive state and a cognitive state.[12] It might be represented as follows:

> *Hybrid Internalism* (*HI*): (**i**) Motivational Internalism. (**ii**) A moral judgment consists in a (**a**) desire and (**b**) a belief.

There are different versions of *HI*, among other things depending on what the object of the desire in (**a**) amounts to: whether it is a single action, a

12 Hybrid internalism entails hybrid expressivism according to which a moral sentence expresses both a non-cognitive state (desire) and a cognitive state (belief). See e.g. Ridge (2006): 302–336; Ridge (2007): 51–76; Ridge (2009): 182–204; Boisvert (2008): 169–203; Boisvert (2014): 22–50, and Hay (2013): 450–474. Cf. Eriksson (2009): 8–35. For overviews of different versions of hybrid expressivism, see Fletcher and Ridge (2014): viii–xvi, and Strandberg (2015b): 91–111. For critical assessments, see e.g. Schroeder (2009): 257–209; Schroeder (2010): Ch. 10, and Strandberg (2015b): 91–111.

certain feature, or all actions that have a certain feature.[13] In what follows, I will consider the last version which is the most common. It can be formulated thus:[14]

> *Action Type Hybrid Internalism (ATHI):* (**i**) Motivational Internalism. (**ii**) The sentence 'It is morally wrong to ϕ' expresses a moral judgment which consists in (**a**) a desire that actions which have a certain feature F are not performed and (**b**) a belief to the effect that ϕing has F.

We can now see that also *ATHI* is subject to the F-G problem. First, consider (**1**). According to this view, (**1**) expresses a moral judgment consisting in a general desire that actions which have a certain feature F are not performed and a belief that lying has F. It follows that a person who accepts (**1**) is motivated to see to it that lying is not performed. Second, consider (**2**). A person who accepts (**2**) need not be motivated to see to it that lying is not performed. Thus, advocates of *ATHI* owe us an explanation of how (**1**) can have the same meaning when it occurs freestanding, as in (**1**), and when it occurs embedded, as in (**2**). As there are no relevant differences between various versions of *HI* that would affect how they fare with regard to the F-G problem, it is plausible to think that it applies to this view in general.[15]

---

13 These alternatives correspond to different versions of hybrid expressivism. See Strandberg (2015b): 91–111.

14 The reason why it is most common is that entails a version of hybrid expressivism which is thought to be able to explain how moral sentences can figure in logically valid arguments. The idea is that irrespective of whether a moral sentence occurs freestanding, such as (**1**), or occurs embedded in a complex sentence, such as (**2**), it expresses a general desire that every action that has a certain feature F is not performed. As every occurrence of (**1**) expresses the very same desire and the relevant belief or proposition, it is argued that an argument like (**1**)–(**3**) is logically valid. See e.g. Ridge (2006): 302–336; Boisvert (2008): 169–203, and Schroeder (2010): Ch. 10. The view is sometimes defended by making an analogy between moral sentences and slurs. It might appear that a slur like 'wop' expresses a negative attitude irrespective of whether it occurs freestanding or embedded. For criticism, see e.g. Strandberg (2015b): 96–104.

15 In defence of *ATHI*, it might be objected that I have misconstrued (**b**). It might be argued that a moral sentence should not be understood to express the *belief* that ϕing has F, but rather the *proposition* that ϕing has F:

Action Type Hybrid Internalism (ATHI*):* (**i**) Motivational Internalism. (**ii**) The sentence 'It is morally wrong to ϕ', expresses a moral judgment which consists in (**a**) a desire that actions which have F are not performed and (**b**) a proposition to the effect that ϕing has F.

According to this view, a person who accepts (**2**) need not believe that lying has F. As a result, she need not be motivated to see to it that lying is not performed. However, there are reasons to think that the revision would not help the view under consideration. First, it might be argued that *ATHI** suffers from other problems than *ATHI*. For example, on *ATHI** it becomes mysterious what it means that a sentence expresses something. According to *ATHI*, a moral sentence expresses mental states, but according to *ATHI** it expresses both a mental state and a proposition. It might be doubted that that there is a plausible notion of 'express' according to which a single sentence can express two types of items that are inherently distinct in the way mental states and propositions

We have seen that the F-G problem applies to *HI* according to which a moral sentence expresses a moral judgment in the form of a complex mental state consisting in a non-cognitive state, in the form of a desire, and a cognitive state, in the form of a belief. Importantly, it is not the claim that a moral sentence expresses a particular type of mental state which makes it subject to the F-G problem. It is rather the claim that a moral sentence expresses a mental state which is motivating that makes is susceptible to this difficulty. Thus, the fact that *HI* is subject to the F-G problem reinforces the suggestion in Section **3**.

More importantly, the fact that *HI* is subject to the F-G problem provides reasons to think that it can be generalized to other versions of internalism in two directions.

First, the F-G problem applies to *HI* according to which a moral sentence expresses a moral judgement consisting in a complex mental state constituted by a desire and a belief. This complex mental state has two significant features: It functions to motivate to action, in virtue of involving a desire, and it functions to represent a certain state of affairs, in virtue of involving a belief. It follows that the F-G problem applies to other views according to which a moral sentence expresses a mental state that has the same characteristics as the mentioned complex mental state. That is, it applies to views according to which a moral sentence expresses a mental state that both functions to motivate and to represent.

Second, the F-G problem applies to *HI* according to which a moral sentence expresses a moral judgment that partly consists in a cognitive state in the form of a belief. According to *HI*, moral sentences might consequently be true or false.[16] This means that the F-G problem might apply to a metaethical view even if it entails that moral sentences have truth-values. Furthermore, it raises the question whether the F-G problem might apply to a view according to which a moral sentence only expresses a belief, provided it has the relevant connection to motivation.

---

are. Moreover, it is difficult to make sense of the notion that a moral judgment consists in a mental state and a proposition. The same type of problems occurs if the view is formulated in terms of the contents of sentences rather than what they express. Second, I do not think that moving from *ATHI* to *ATHI** makes any important difference to the argument above. According to *ATHI**, (**1**) expresses a desire that actions having F are not performed. However, it is unclear why this would be the case when (**1**) occurs embedded in a complex sentence such as (**2**). To illustrate, consider a moral sceptic who accepts (**2**). Assume that she denies that there are any actions which are wrong because she denies that there are any actions which have F. It is difficult to understand why she would need to have a desire that actions having F are not performed and be accordingly motivated. In Strandberg (2015b): 99–102, I argue that the most influential version of hybrid expressivism suffers from a similar problem.

16 Ridge distinguishes between 'ecumenical expressivism' and 'cognitivist expressivism'. On the former view, a moral sentence expresses both a belief and a desire, but it is not the case that the sentence is true if the belief is true. On the latter view, a moral sentence expresses both a belief and a desire, and the sentence *is* true if the belief is true. Ridge (2006): 302–336. Cf. Barker (2000): 268–279, and Boisvert (2008): 169–203.

5.3. *Sui Generis* Internalism

We should next briefly consider the third version of state internalism:

> *Sui Generis Internalism (SGI):* (**i**) Motivational Internalism. (**ii**) A moral judgment consists in a *sui generis* mental state ('besire').[17]

The *sui generis* mental state to which the view refers has two aspects: First, it represents a certain state of affairs as being the case. Second, it motivates to action. Thus, 'It is wrong to ϕ' expresses a moral judgment in the form of a *sui generis* state which represents it as being the case that ϕing is wrong and motivates to see to it that ϕing is not performed. A person who is in this *sui generis* state is consequently motivated to see to it that ϕ is not performed.

We have already seen why *SGI* is subject to the F-G problem.[18] As noted in the last section, the problem applies to views according to which a moral sentence expresses a moral judgment consisting in a mental state that both functions to motivate and represent. Hence, *SGI* is subject to the F-G problem for the same reason as hybrid internalism (*HI*).

6. Object Internalism and the Frege-Geach Problem

In this section, I will argue for the controversial claim that object internalism is subject to the F-G problem.

6.1. Cognitivist Internalism

We saw in Section **4** that object internalism is most plausibly combined with the view that moral judgments consist in beliefs. We get:

> *Cognitivist Internalism* (*CI*): (**i**) Motivational Internalism. (**ii**) A moral judgment consists in a belief.

The view can be formulated both in terms of what a moral sentence expresses and in terms of its content. Formulated in the first manner: The sentence 'It is wrong to ϕ' expresses a moral judgment in the form of a moral belief that ϕing is wrong. If a person holds this belief, she is motivated to see to it that ϕing is not performed. Formulated in the second manner: The content of

---

17 For clarifications of the nature of this kind of mental state, see e.g. Millikan (1995): 185–200. Cf. Zangwill (2008): 50–59. The term 'besire' was coined in Altham (1986): 284. It is not always clear whether a particular author advocates *sui generis* internalism (*GCI*) or cognitivist internalism (*CI*). However, among the authors that can be interpreted to defend the former view, see e.g. Little (1997): 59–79; Bedke (2009): 189–209, and Swartzer (2019): 975–988.

18 I discuss this view more thoroughly in Strandberg (2015a): 1–15. Cf. Björnsson (2001): 87.

the sentence 'It is wrong to ϕ' consists in the proposition: it is wrong to ϕ. A belief that has such a proposition as its object is motivating in the indicated manner.

Let us start by clarifying *CI*. It might be asked what explains that a moral belief is motivating on this view. The most plausible answer seems to be: It involves a moral proposition as its object.[19] There are mainly two reasons for this contention. First, it is generally accepted that not all beliefs are motivating.[20] It must consequently be something about moral beliefs or, more broadly, normative beliefs, which explains why they, in contrast to other beliefs, are motivating. The only thing that seems to distinguish them from other beliefs is that they involve a certain proposition as their object. This view finds evidence in various claims by cognitivist internalists and other philosophers who have commented on this view.[21] Second, in order for moral beliefs that are motivating to be genuine *beliefs*, and not some other type of mental states, it needs to be a moral proposition which explains why they are motivating. Assume that it is denied that it is moral propositions that make moral beliefs motivating. It has then to be something about the nature of the kind of mental states that make up moral beliefs which explains that they are motivating. In that case, it is difficult to see that these mental states are genuine beliefs rather than desires, or beliefs in conjunction with desires, or some other type of mental states, like *sui generis* mental states. Moreover, on this assumption *CI* would not be an instance of object internalism, but of state internalism, in which case it can be argued that it is vulnerable to the arguments above. It is noteworthy that none of these arguments appealed specifically to the claim that moral sentences express desires, but to the claim that they express states belonging to a kind of mental states that is motivating. If *CI* is assumed to state that it is something about the nature of the kind of mental states constituting moral beliefs which explains that they are motivating, it might in other words be suspected that these arguments can be directed against this view as well.

---

19 A moral proposition which explains motivation need not be moral in the sense that it explicitly contains a 'thin' moral concept such as wrongness. See e.g. McDowell (1979): 14.

20 But see Bromwich (2010): 343–367.

21 For explanations of the nature of this kind of beliefs, see e.g. Noggle (1997): 90–91; Jacobson-Horowitz (2006): 561–580, and Pearson (2015): 255–276. See also e.g. Lewis (1988): 323–332; Wedgwood (1995): 273–288; C. Miller (2008): 222–266, and Tanyi (2014): 331–348. Among the authors that can be interpreted to advocate this view, see e.g. Nagel (1970), Part Two; McDowell (1978): 13–29; McDowell (1979): 331–350; Platts (1979): 255–263; McNaughton (1988): Ch. 7; Wiggins (1991): 51–85; Dancy (1993): Ch. 2; Dancy (1999): 217–223; van Roojen (2002): 26–49; Tenenbaum (2006): 235–264, and Bromwich (2010): 343–367. See also Mele (1996): 747–753; Scanlon (1998): 37–41, and Shafer-Landau (2003): Ch. 5.

Thus, according to *CI* a moral belief is motivating in virtue of involving a moral proposition as its object, *not* in virtue of being a belief. The motivating force of a moral belief does not depend on the fact that one *believes* so and so, but on *what* one believes: the moral proposition constituting the object of the belief. In this way, a moral proposition can be said to bestow motivating force on a belief, whereas the belief does not have any motivating force merely in virtue of being a belief.[22] We might put it in the following way: Motivation is *inherent* to moral propositions in the sense that what explains that a moral belief is motivating is that a moral proposition, in virtue of its nature, is such as to make beliefs motivating. However, motivation is *not inherent* to beliefs since they are not, in virtue of their nature, motivating. In other words, an explanation of why a moral belief is motivating refers to a feature the moral proposition has in virtue of being a moral proposition, not to a feature the belief has in virtue of being a belief.

According to *CI*, a moral proposition consequently has two aspects. First, it has a *cognitive aspect* in that it, like other propositions, represents a certain state of affair.[23] If the proposition is the object of a belief, it is presented as being true or, to put it in another way, the state of affair in question is presented as being the case. Second, it has a *motivational aspect* in that it, *unlike* other propositions, is such as to make beliefs motivating. Moreover, *both* these aspects are *inherent* to a moral proposition in the sense just mentioned. Thus, the fact that a moral proposition represents a certain state of affairs is explained by the nature of the proposition, not by being an object of a certain belief. Likewise, the fact that a belief which involves a moral proposition as its object is motivating is explained by the nature of the proposition, not in virtue of being a belief.

There are some issues with regard to *CI* that should be mentioned but that are not pertinent to the present discussion. One issue concerns how the claim that a belief is motivating should be spelled out. According to an influential view, the difference between beliefs and desires is a matter of 'directions of fit': Beliefs aim at fitting the world whereas desires aim at getting the world fitting them, which in turn can be accounted for in different ways. In a similar vein, *CI* can be understood to imply that moral beliefs, in

22 Hilla Jacobson-Horowitz aptly puts the view as follows: 'In the sense relevant to their role in practical reasoning, then, it is not the psychological mode of beliefs which determines their dominant direction of fit and thus their motivation character (in this respect their mode is "transparent"), but rather their content. Thus, if a belief's content is a normative, requiring, content—as is the case with moral beliefs—the belief has a requiring character and may play a motivational role. The content of a belief being a normative content endows it with requiring character, its psychological mode—which is responsible for its classification as a cognitive attitude—notwithstanding' (Jacobson-Horowitz (2006): 563). In the same vein, Ralph Wedgwood writes that in case there are beliefs that are motivating, 'they would have this tendency in virtue of their *content*, not simply in virtue of being beliefs' (Wedgwood (1995): 274). See also references to Noggle and Pearson above.

23 Cf. Wedgwood (2007): 59.

contrast to other beliefs, have both these aims in virtue of involving a moral proposition.[24] Another issue concerns the connection between beliefs and desires.[25] On one view, a moral belief is motivating in the sense that what motivates is the belief itself without the help of any desire. On another view, a moral belief is motivating in the sense that it by itself gives rise to a desire.[26] However, on *CI*, what makes the belief motivating on either alternative would be a moral proposition.[27] [28]

Let us now return to the F-G problem. Think again of a freestanding moral sentence, such as (**1**). If we grant cognitivism, *CI* might seem plausible since it is reasonable to assume that a person who asserts the sentence is motivated to see to it that lying is not performed. However, a person might assert a sentence where this sentence occurs embedded, such as (**2**), without being thus motivated.

According to *CI*, (**1**) expresses a belief that has the following proposition as its object: *it is wrong to lie*. I will refer to this as '*the first belief*'. The sentence (**2**) expresses a belief that has a proposition as its object where this consists in a conditional proposition: *if it is wrong to lie, then it is wrong to get one's little brother to lie*. I will refer to this as '*the second belief*'. Thus, (**1**) expresses a belief that has as its object a certain moral proposition and (**2**) expresses a belief that has as its object a proposition where this moral proposition constitutes the antecedent. The moral proposition in question is: *it is wrong to lie*.

Now, it can be argued that advocates of *CI* need to explain how it can be the *same proposition* in these two cases. We saw above that on this view a moral proposition has two aspects: a cognitive and a motivational aspect. We also saw that a moral proposition has both these aspects in virtue of being a certain proposition and not in virtue of being an object of a particular belief. When it comes to the *cognitive aspect*, the proposition in question is clearly the same with regard to the two beliefs that are expressed in (**1**) and (**2**): In both these beliefs, the proposition represents a state of affairs, viz. that it is wrong to lie. However, when it comes to the *motivational aspect*, it might be

24 Cf. Wedgwood (1995): 274, and Jacobson-Horowitz (2006): 566.

25 See e.g. Shafer-Landau (2003): 122–123, and Persson (2005): 54.

26 The first alternative seems to be adopted by e.g. McDowell and the second by e.g. Nagel.

27 Further, it might be asked how the contention that a moral proposition has this motivational aspect should be understood. In an early paper, Ralph Wedgwood argues that there is no plausible conception of propositions which is compatible with the claim that propositions make beliefs motivating (Wedgwood (1995): 273–288). In what follows, I will for sake of the argument grant that propositions can have this aspect.

28 The version of internalism under consideration is often conjoined with a denial of the Humean theory of motivation. However, as *CI* is understood here, this inference is not obvious. According to *CI*, a moral belief is motivating, but it seems at least conceivable that this belief is caused by a desire and that no belief is motivating unless it is caused by a desire. This view is compatible with the Humean contention that no belief is sufficient for itself for motivation but that all motivation requires an independently existing desire.

asked how it can be the same proposition with regard to the two beliefs that are expressed in (**1**) and (**2**). According to *CI*, a moral belief is motivating in virtue of involving a moral proposition as its object, not in virtue of being a belief. However, the second belief, the belief expressed in (**2**), is clearly not motivating in spite of having as its object a proposition of which this moral proposition constitutes a part. It thus seems that the moral proposition in question makes the first belief motivating whereas it does not make the second belief motivating. Moreover, the proposition does not seem to affect the motivating force of the second belief in any respect whatsoever. In other words, it appears that in these two beliefs the moral proposition remains constant as regards its cognitive aspect but not as regards its motivational aspect. This makes it justified to ask how it can be the *same proposition*.

It might be responded that *CI* has the resources to avoid the F-G problem. The reason why non-cognitivist internalism (*NCI*) and hybrid internalism (*HI*) are subject to this problem is that they entail that a moral sentence expresses a mental state which is motivating. Likewise, they entail that the content of a moral sentence consists in a mental state which is motivating. However, *CI* need not understand the meaning of moral sentences in terms of mental states. In particular, it does not claim that the content of a moral sentence is constituted by a mental state, but a proposition, and so it might seem that the F-G problem does not apply to this view.

However, this response is misguided since the F-G problem for *CI* can be formulated in terms of the content of moral sentences. The content of (**1**) is the proposition: it is wrong to lie. The content of (**2**) is the proposition: if it is wrong to lie, then it is wrong to get one's little brother to lie. Thus, the proposition that is the content of (**1**) constitutes the antecedent of the content of (**2**): *it is wrong to lie*. When it comes to the cognitive aspect, this proposition is clearly the same with regard to both (**1**) and (**2**): In both cases, it represents the same state of affairs. However, when it comes to the motivational aspect it might be asked how it can be the same proposition in the two cases. The second belief, the belief which has as its object the content of (**2**), is clearly not motivating, despite the fact that it has as its object a proposition of which the moral proposition under consideration is a part. Again, it might be asked how it can be the *same proposition*.

We are now in the position to strengthen the formulation of the F-G problem for *CI*. As we saw earlier, according to this view a moral proposition has both the cognitive and the motivational aspect in virtue of being a particular proposition. In other words, both aspects of a moral proposition are *inherent* to it in the sense that it is in virtue of its nature that a moral proposition has these aspects. In view of this fact, it is especially worrying that the proposition appears to have both aspects with regard to the first belief, but only one aspect with regard to the second belief, since the nature of a proposition cannot be affected by being combined with another proposition.

The F-G problem for *CI*, understood in terms of the content of moral sentences, thus amounts to this. According to this view, the content of a freestanding sentence, such as (**1**), consists in a proposition that has two aspects: a cognitive and a motivational aspect. However, when the sentence occurs embedded in a complex sentence, such as (**2**), its content consists in a proposition that seems to have the first aspect but not the second. Therefore, advocates of *CI* owe us an explanation of how a freestanding and an embedded occurrence of a moral sentence can have the same proposition as their content. As a consequence, they owe us an explanation of how a freestanding and an embedded moral sentence can have the same meaning.

We can now see that the F-G problem for *CI* also can be formulated in terms of the logical validity of moral arguments. Recall the argument (**1**)–(**3**). If the antecedent of (**2**) does not have the same meaning as (**1**), (**3**) would not follow. As (**3**) does follow, they have to have the same meaning. However, we have seen that there are reasons to doubt that (**1**) and the antecedent of (**2**) can have the same meaning according to *CI*.

6.2. Two Defences of Cognitivist Internalism Considered

In order to evade the F-G problem, defenders of *CI* need to explain how it can be the same moral proposition with regard to the two beliefs we considered above, in spite of the fact that the proposition has the cognitive aspect in both cases but appears to lack the motivational aspect in the latter case.

According to the first defence, a moral proposition that occurs separately, without being combined with another proposition, has both the cognitive and the motivational aspect. However, a moral proposition that is combined with another proposition has the cognitive aspect but lacks the motivational aspect.

It is not difficult to see that this defence is misguided. A proposition does not change by being combined with other propositions. It thus contributes in the same way irrespective of whether it is occurs alone or as an antecedent of a conditional. It is then difficult to see how this suggestion can explain why a moral proposition has the motivational aspect in one case but not in the other. Moreover, advocates of *CI* cannot uphold the idea that it is only when a moral proposition occurs separately that it has the motivational aspect. Think of a belief that has as its object the proposition: it is wrong to lie *and* it is wrong to get one's little brother to lie. It is reasonable to think that, on *CI*, a person who holds this belief is motivated.

According to the second defence, a moral proposition that is *believed* has both the cognitive and the motivational aspect. However, a moral proposition that is *not* believed has the former aspect but lacks the latter. This presumably appears as the most plausible defence, but as we will see, it suffers from basically the same difficulty as the first response.

There are different views about what it means that a proposition is believed, but in the present context these differences are not essential, and it

should not be difficult to translate what I say to the preferred vocabulary. We may consequently describe it in the following commonsensical way. Consider the first belief, which has as its object the proposition: it is wrong to lie. In this case, the proposition is believed because the belief presents it as being true. Formulated in another way, the proposition is believed because the belief presents the state of affairs in question as being the case. Consider next the second belief, which has as its object the proposition: if it is wrong to lie, it is wrong to get one's little brother to lie. In this case, the moral proposition at issue—it is wrong to lie—is not believed, because the belief has a conditional proposition as its object where this proposition constitutes the antecedent. According to the second defence, it is the different relations these two beliefs have to this moral proposition which explains that it has the motivational aspect in the first case but lacks it in the second case.

We can now see that the second defence fails for the same reason as the first one. According to the present defence, the reason why the first belief is motivating whereas the second belief is not, is that the moral proposition in question is believed in the first case but not the other. We have already seen that whether a proposition is believed or not depends on whether it is presented as being true or is part of a complex proposition such that the belief does not present the proposition as being true. As regards the first belief, the moral proposition is presented as being true. As regards the second belief, the moral proposition is not presented as being true because it constitutes the antecedent of a conditional. However, we have already seen that a proposition is not affected in any way by being combined with another proposition so as to become part of a complex proposition, such as the antecedent of a conditional. Whether a proposition is believed or not cannot influence the nature of the proposition. Hence, the present suggestion is unable to explain why the moral proposition has the motivational aspect with regard to the first belief but lacks it with regard to the second belief.

There is also another difficulty for the second defence. According to *CI*, a moral proposition has both a cognitive and a motivational aspect, and it has both these aspects in virtue of being a certain type of proposition. The cognitive aspect of the moral proposition under consideration is clearly the same with regard to both the beliefs we have considered: It represents a certain state of affairs. The fact that the first belief represents it as true whereas the second belief does not, cannot alter this fact. In both beliefs, the contribution this proposition makes is consequently the same as far as the cognitive aspect in concerned. Now, since the motivational aspect of the moral proposition also is supposed to be a feature it has in virtue of being a certain proposition, it seems that the same consideration should apply to it too. It consequently seems that the proposition should contribute in the same way as regards the two beliefs when it comes to the motivational aspect respect as well. But it does not. As a result, the second defence cannot help to explain why it is a matter of the same proposition in the two cases.

7. Two Versions of Conditional Internalism

In section **4**, I distinguished between unconditional and conditional versions of internalism. On the first view, the necessary connection between moral judgments and motivation holds for every person, whereas it on the second version holds only for those who satisfy a certain condition. The distinction cuts through the division between state and object internalism, which means that there are unconditional and conditional versions of all four types of internalism I have considered. In subsequent sections, I argued that various unconditional versions of internalism are subject to the F-G problem. In this section, I will consider whether conditional versions of internalism is able to avoid it.[29] Consider:

> *Conditional Internalism*: It is conceptually necessary that if a person judges that it is morally wrong to ϕ, then she is, at least to some extent, motivated to see to it that ϕing is not performed, given that she satisfies condition C.

In this claim, 'C' can be specified in a number of different ways, but it has to be such that it does not render the internalist claim trivially true. What is important for our purposes, however, is that there are two broad but distinct kinds of conditional internalism.

According to *strong conditional internalism*, there are cases where a person's judgment to the effect that it is wrong to ϕ is sufficient by itself for her to be motivated to see to it that ϕing is not performed. We might test whether a particular conditional internalist claim is of this kind by considering whether there is any possible world where a person's moral judgment is sufficient by itself for her to be accordingly motivated. If there *is* such a possible world, the claim in question belongs to this kind. According to this view, C can be understood to specify the *absence* of a hindrance of some sort for the judgment to be motivating, such as absence of certain mental conditions or 'non-normal' circumstances.[30] In case the hindrance in question is absent, the moral judgment is sufficient by itself for motivation.

---

29 For defences of unconditional internalism, see e.g. Lenman (1999): 441–457; Joyce (2001): 17–29, and Bromwich (2016): 452–471. McDowell's version of cognitivist internalism is presumably an instance of unconditional internalism, since he maintains that a person who is not accordingly motivated does not hold the moral belief in question. See e.g. McDowell (1979): 16. Cf. McNaughton (1988): Ch. 8.

30 Unfortunately, it is not always entirely clear whether a certain version of conditional internalism should be classified as strong or weak. However, in strong conditional internalism condition C seems often to be understood as the absence of particular mental conditions, such as addiction, apathy, compulsion, emotional disturbance, etc. See e.g. Dancy (1993): 25, and Svavarsdóttir (1999): 165. (However, Svavarsdóttir does not defend this view.) Alternatively, it might be understood as the absence of 'non-normal' circumstances. See e.g. Blackburn (1998): 59–68; Gibbard (2003): 152–154, and Dreier (1990): 9–14. For criticism, see e.g. Strandberg (2012): 81–91.

Now, strong conditional internalism entails that it is something about the very moral judgment which explains that it can be sufficient all by itself for motivation. In line with the distinction between state and object internalism, there seems to be two alternatives: Either the moral judgment involves a mental state which belongs to a kind of mental states that is motivating, *or* it has a moral proposition as its object which makes it motivating. As a consequence, the F-G problem applies to strong conditional internalism in the same manner as it applies to unconditional internalism.

According to *weak conditional internalism*, there are *no* cases where a person's judgment that it is wrong to ϕ is sufficient by itself for her to be motivated to see to it that ϕing is not performed. Again, we might test whether a certain internalist claim is of this kind by considering whether there is any possible world where a person's moral judgment is sufficient by itself for motivation. If there is *no* such possible world, the claim in question belongs to this kind. According to this view, C can be understood to specify something that *needs* be present in order to assure that a person who makes a moral judgment is motivated.[31] Thus, a person's moral judgment is not such that it all by itself can be sufficient to explain her motivation. Rather, it is her moral judgment *in conjunction with* the fact that she satisfies C that provides such an explanation. As a consequence, it is difficult to see that the F-G problem applies to weak conditional internalism on either of the two lines I developed above. The most prevalent version of this view understands C in terms of practical rationality.[32]

Thus, although there are versions of conditional internalism that escape the F-G problem, this is by no means the case as regards every instance of this view. The versions of internalism that are subject to this problem has the following in common: They entail that a person's judgment that it is wrong to ϕ can be sufficient by itself for her to be motivated to see to it that ϕing is not performed. Accordingly, all types of unconditional internalism and strong conditional internalism are subject to the F-G problem, whereas weak conditional internalism is not. It should be stressed that this does not mean that the arguments of the previous sections are insignificant. It is widely thought that conditional versions of internalism are problematic for various reasons. Especially, it has been shown difficult to come up with a notion of practical rationality that does not threaten to make the resulting claims vacuous.[33] Moreover, the only versions of hybrid internalism (*HI*) I know of are instances of unconditional internalism, and most versions of cognitivist internalism (*CI*) appear to be instances of unconditional internalism or strong conditional internalism.

---

31 See e.g. Korsgaard (1996): 5–25; Smith (1994): Ch. 3; Wedgwood (2007): Ch. 1, and van Roojen (2010): 495–525. For criticism, see e.g. Strandberg (2013): 25–51.

32 See Smith (1994), esp. Ch. 3. Cf. Korsgaard (1986): 5–25, and Wedgwood (2007): Ch. 1.

33 See e.g. Lenman (1996): 298–299; Sayre-McCord (1997): 64–65; Svavarsdóttir (1999): 164–165; A. Miller (2003): 221; Roskies (2003): 53, and Schroeter (2005): 4.

8. Three Metaethical Lessons

In this paper, I have argued that the Frege-Geach problem applies to the two basic forms of internalism: state internalism and object internalism. It applies to state internalism in all its three versions: non-cognitivist, hybrid, and *sui generis* internalism. Moreover, I have maintained that it also applies to object internalism in the form of cognitivist internalism. However, I also pointed out that the F-G problem does not apply to weak conditional versions of internalism. I conclude the paper by drawing three general lessons concerning the scope of the Frege-Geach problem.

First, the Frege-Geach problem might apply to a metaethical view irrespective of what type of mental state a moral sentence is claimed to express. According to non-cognitivist internalism, a moral sentence expresses a non-cognitive state, but according to hybrid internalism it expresses a non-cognitive state in combination with a cognitive state, and according to *sui generis* internalism it expresses a distinct type of mental state. Further, according to cognitivist internalism a moral sentence expresses a purely cognitive state. However, all these views are subject to the Frege-Geach problem.

Second, the Frege-Geach problem might apply to a metaethical view even if it entails that moral sentences can be true or false. According to hybrid internalism, *sui generis* internalism, and cognitivist internalism, moral sentences have truth-values. However, they are still subject to this problem.

Third, the Frege-Geach problem might apply to a metaethical view even if it emphasizes that the content of a moral sentence consists in a proposition. Assume that it is argued that the content of a sentence cannot consist in a mental state, like a belief or desire, but must consist in a proposition, or some other abstract entity. However, we have seen that the problem applies to cognitivist internalism even if this view is understood to claim that the content of a moral sentence consists in a moral proposition.

References

Altham, J.E.J. (1986), 'The Legacy of Emotivism', In *Fact, Science and Morality*, eds. G. McDonald and C. Wright, Oxford: Blackwell, pp. 275–288.

Barker, S. (2000), 'Is Value Content a Component of Conventional Implicature?', *Analysis*, Vol. 60, pp. 268–279.

Bedke, M.S. (2009), 'Moral Judgments Purposivism: Saving Internalism from Amoralism', *Philosophical Studies*, Vol. 144, pp. 189–209.

Björnsson, G. (2001), 'Why Emotivists Love Inconsistency', *Philosophical Studies*, Vol. 167, pp. 81–108.

Björnsson, G. et al. (2015), 'Motivational Internalism: Contemporary Debates', In *Motivational Internalism*, eds. G. Björnsson, C. Strandberg, R. Francén, J. Eriksson and F. Björklund, Oxford: Oxford University Press, 2015, pp. 1–20.

Blackburn, S. (1998), *Ruling Passions*, Oxford: Clarendon Press.

Blome-Tillmann, M. (2009), 'Moral Non-cognitivism and the Grammar of Morality', *Proceedings of the Aristotelian Society*, Vol. CIX, pp. 279–309.

Boisvert, D. (2008), 'Expressive-Assertivism', *Pacific Philosophical Quarterly*, Vol. 89, pp. 169–203.

Boisvert, D. (2014), Expressivism, Nondeclaratives, and Success-conditional Semantics, In *Having it Both Ways*, eds. G. Fletcher and M. Ridge, Oxford: Oxford University Press, pp. 22–50.

Bromwich, D. (2010), 'Clearing Conceptual Space for Cognitivist Motivational Internalism', *Philosophical Studies*, Vol. 148, pp. 343–367.

Bromwich, D. (2016), 'Motivational Internalism and the Challenge of Amoralism', *European Journal of Philosophy*, Vol. 24, pp. 452–471.

Cholbi, M. (2006). 'Belief attribution and the Falsification of Motive Internalism', *Philosophical Psychology*, Vol. 19, pp. 607–616.

Cuneo, Terence (1999), 'An Externalist Solution to the "Moral Problem"', *Philosophy and Phenomenological Research*, Vol. LIX, pp. 359–380.

Dancy, J. (1993), *Moral Reasons*, Oxford: Blackwell.

Dancy, J. (1999), 'Motivation, Dispositions and Aims', *Theoria*, Vol. LXV, pp. 212–223.

Dreier, J. (1990), 'Internalism and Speaker Relativism', *Ethics*, Vol. 101, pp. 6–26.

Eklund, M. (2009), 'The Frege-Geach Problem and Kalderon's Moral Fictionalism', *Philosophical Quarterly*, Vol. 58, pp. 705–712.

Eriksson, J. (2009), 'Homage to Hare: Ecumenism and the Frege-Geach Problem', *Ethics*, Vol. 120, pp. 8–35.

Fletcher, G. and M. Ridge (2014), 'Introduction', In *Having it Both Ways*, eds. G. Fletcher and M. Ridge, Oxford: Oxford University Press, pp. viii–xvi.

Francén, R. (2010), 'Moral Motivation Pluralism', *The Journal of Ethics*, Vol. 14, pp. 117–148.

Geach, P. (1960), 'Ascriptivism', *Philosophical Review*, Vol. 69, pp. 221–225

Geach, P. (1965), 'Assertion', *Philosophical Review*, Vol. 74, pp. 449–465.

Gert, J. and A.R. Mele (2005), 'Lenman on Externalism and Amoralism: An Interplanetary Exploration', *Philosophia*, Vol. 32, pp. 275–283.

Gibbard, A. (2003), *Thinking How to Live*, Cambridge, Mass.: Harvard University Press.

Hay, R. (2013), 'Hybrid Expressivism and the Analogy between Pejoratives and Moral Language', *European Journal of Philosophy*, Vol. 21, pp. 450–474.

Jacobson-Horowitz, H. (2006), 'Motivational Cognitivism and the Argument from Direction of Fit', *Philosophical Studies*, Vol. 127, pp. 561–580.

Joyce, R. (2001), *The Myth of Morality*, Cambridge: Cambridge University Press.

Kalderon, M.E. (2005), *Moral Fictionalism*, Oxford: Clarendon.

Kauppinen, A. (2008), 'Moral Internalism and the Brain', *Social Theory and Practice*, Vol. 34, pp. 1–24.

Korsgaard, C. (1986), 'Skepticism About Practical Reason', *Journal of Philosophy*, Vol. XXXIII, pp. 5–25.

Lenman, J. (1996), 'Belief, Desire and Motivation: An Essay in Quasi-hydraulics', *American Philosophical Quarterly*, Vol. 33, pp. 291–301.

Lenman, J. (1999), 'The Externalist and the Amoralist', *Philosophia*, Vol. 27, pp. 441–457.

Lewis, D. (1988), 'Desire as Belief', *Mind*, Vol. XCVII, pp. 323–332.

Lillehammer, H. (2002), 'Moral Cognitivism', *Philosophical Papers*, Vol. 31, pp. 1–25.

Lippert-Rasmussen, K. (2002), 'Must Morality Motivate?', *Danish Yearbook of Philosophy*, Vol. 37, pp. 7–36.

Little, M.O. (1997), 'Virtue as Knowledge: Objections from the Philosophy of Mind, *Noûs*, Vol. 31, pp. 59–79.

McDowell, J. (1978), 'Are Moral Requirements Hypothetical Imperatives?', *The Aristotelian Society*, Supplementary Vol. LII, pp. 13–29.

McDowell, J. (1979), 'Virtue and Reason', *Monist*, Vol. 62, pp. 331–350.

McNaughton, D. (1988), *Moral Vision*, Oxford: Blackwell.

Mele, A.R. (1996), 'Internalist Moral Cognitivism and Listlessness', *Ethics*, Vol. 106, pp. 727–753.

Milevski, V. (2015), 'The Argument from Moral Psychology', *Belgrade Philosophical Annual*, Vol. XXVIII, pp. 113–126.

Miller, A. (2003), *An Introduction to Contemporary Metaethics*, Cambridge: Polity Press.

Miller, C. (2008), ''Motivation in Agents', *Noûs*, Vol. 42, pp. 222–266.

Miller, C.B. (2008), 'Motivational Internalism', *Philosophical Studies*, Vol. 139, pp. 233–255.

Millikan, R.G. (1995), 'Pushmi-Pullyu Representations', *Philosophical Perspectives*, Vol. 9, pp. 186–200.

Nagel, T. (1970), *The Possibility of Altruism*, New York: Oxford University Press.

Noggle, R. (1997), 'The Nature of Motivation (and Why it Matters Less to Ethics than One Might Think)', *Philosophical Studies*, Vol. 87, pp. 87–111.

Pearson, G. (2015), 'What are Sources of Motivation?', *Proceedings of the Aristotelian Society*, Vol. CXV, pp. 255–276.

Persson, I. (2005), *The Retreat of Reason*, Oxford: Oxford University Press.

Platts, M. de Bretton (1979), *Ways of Meaning*, London: Routledge.

Ridge, M. (2003), 'Non-cognitivist Pragmatics and Stevenson's "Do so as Well!"' *Canadian Journal of Philosophy*, Vol. 4, pp. 563–574.

Ridge, M. (2006), 'Hybrid Expressivism: Finessing Frege', *Ethics*, Vol. 116, pp. 302–336.

Ridge, M. (2007), 'Hybrid Expressivism: The Best of Both Worlds?', In *Oxford Studies in Metaethics*, Vol. 2, ed. Russ Shafer-Landau, Oxford: Oxford University Press, pp. 51–76.

Ridge, M. (2009), 'Moral Assertion for Expressivists', *Philosophical Issues*, Vol. 19, pp. 182–204.

Roojen, M. van (2002), 'Humean and Anti-Humean Internalism about Moral Judgements', *Philosophy and Phenomenological Research*, Vol. LXV, pp. 26–49.

Roojen, M. van (2010), 'Moral Rationalism and rational Amoralism', *Ethics*, Vol. 120, pp. 495–525.

Roskies, A. (2003), 'Are Ethical Judgments Intrinsically Motivational? Lessons from "Acquired Sociopathy"', *Philosophical Psychology*, Vol. 16, pp. 51–66.

Sayre-McCord, G. (1997), 'The Metaethical Problem', *Ethics*, Vol. 108, pp. 55–83.

Scanlon, T.M. (1998), *What We Owe to Each Other*, Cambridge, Mass.: Harvard University Press.

Schroeder, M. (2008), *Being For*, Oxford: Clarendon.

Schroeder, M. (2009), 'Hybrid Expressivism: Virtues and Vices', *Ethics*, Vol. 119, pp. 257–309.

Schroeder, M. (2010), *Non-cognitivism in Ethics*, New York: Routledge.

Schroeter, F. (2005), 'Normative Concepts and Motivation', *Philosophers' Imprint*, Vol. 5, pp. 1–23.

Searle, J. (1962), 'Meaning and Speech-Acts', *Philosophical Review*, Vol. 71, pp. 423–432.

Shafer-Landau, R. (2003), *Moral Realism. A Defence*, Oxford: Clarendon Press.

Sinnott-Armstrong, W. (2000), 'Expressivism and Embedding', *Philosophy and Phenomenological Research*, Vol. 61, pp. 677–693.

Smith, M. (1994), *The Moral Problem*, Oxford: Blackwell.

Sneddon, A. (2009) 'Alternative Motivation: A New Challenge to Moral Judgment Internalism', *Philosophical Explanations*, Vol. 12, pp. 41–53.

Strandberg, C. (2011), 'The Pragmatics of Moral Motivation', *The Journal of Ethics*, Vol. 15, pp. 341–369.

Strandberg, C. (2012), 'Expressivism and Dispositional Desires', *American Philosophical Quarterly*, Vol. 49, pp. 81–91.

Strandberg, C. (2013), 'An Internalist Dilemma—and an Externalist Solution', *The Journal of Moral Philosophy*, Vol. 10, pp. 25–51.

Strandberg, C. and F. Björklund (2013), 'Is Moral Internalism Supported by Folk Intuitions?', *Philosophical Psychology*, Vol. 26, pp. 319–335.

Strandberg, C. (2015a), 'Can the Embedding Problem be Generalized?', *Acta Analytica*, Vol. 30, pp. 1–15.

Strandberg, C. (2015b), 'Options for Hybrid Expressivism', *Ethical Theory and Moral Practice*, Vol. 1, pp. 91–111.

Strandberg, C. (2016), 'Aesthetic Internalism and two Normative Puzzles', *Studi di estetica*, Vol. XLIV, pp. 23-70.

Svavarsdóttir, S. (1999), 'Moral Cognitivism and Motivation', *Philosophical Review*, Vol. 108, pp. 161–219.

Swartzer, S. (2013), 'Appetitive Besires and the Fuss about Fit', *Philosophical Studies*, Vol. 165, pp. 975–988.

Tanyi, A. (2014), 'Pure Cognitivism and Beyond', *Acta Analytica*, Vol. 29, pp. 331-348.

Tenenbaum, S. (2006), 'Direction of Fit and Motivational Cognitivism', In *Oxford Studies in Metaethics*, Vol. 1, ed. Russ Shafer-Landau, Oxford: Oxford University Press, pp. 235–264.

Tresan, J. (2006) 'De Dicto Internalist Cognitivism', *Noûs*, Vol. 40, pp. 143–165.

Tresan, J. (2009), 'Metaethical Internalism: Another Neglected Distinction', *Journal of Ethics*, Vol. 13, pp. 51–72.

Wedgwood, R. (1995), 'Theories of Content and Theories of Motivation', *European Journal of Philosophy*, Vol. 3, pp. 273–288.

Wedgwood, R. (2007), *The Nature of Normativity*, Oxford: Oxford University Press.

Wiggins, D. (1991), 'Moral Cognitivism, Moral Relativism and Motivating Moral Beliefs', *Proceedings of the Aristotelian Society*, Vol. 91, pp. 51–85.

Zangwill, N. (2007), 'The Indifference Argument', *Philosophical Studies*, Vol. 138, pp. 91–124.

Zangwill, N. (2008), 'Besires and the Motivation Debate', *Theoria*, Vol. 74, pp. 50–59.

John Eriksson
University of Gothenburg, Sweden

Original Scientific Paper
UDC 17.022.1:165
17.03:161/164

EXPLAINING DISAGREEMENT: CONTEXTUALISM, EXPRESSIVISM AND DISAGREEMENT IN ATTITUDE

Abstract. *A well-known challenge for contextualists is to account for disagreement. Focusing on moral contextualism, this paper examines recent attempts to address this challenge by using the standard expressivist explanation, i.e., explaining disagreement in terms of disagreement in attitude rather than disagreement in belief. Assuming that the moral disagreements can be explained in terms of disagreement in attitude, this may seem as a simple solution for contextualists. However, it turns out to be easier said than done. This paper examines a number of different ways in which disagreement in attitude can be incorporated into a contextualist framework and argues that each suggestion is problematic. In particular, the purported explanations of disagreement fail to adequately explain intuitive occurrences of disagreement, the robustness of disagreement intuitions and/or locate the disagreement in the intuitively right place.*

Keywords: *contextualism; expressivism; disagreement; disagreement in belief; disagreement in attitude*

Introduction

Contextualism is a view according to which the meaning of certain terms is incomplete and fixed by the context of utterance.[1] Indexical terms serve as paradigmatic examples. The meaning of "I," "here" or "now," for example, depend on the context (speaker, place and time respectively). A speaker who says "It is hot here" while in Los Angeles picks out a different place than someone who utters the same words in Alaska. Contextualism is also a semantic doctrine that many philosophers find plausible in other domains, e.g., taste, aesthetics and morality. The idea is that terms in these domains, for example, "delicious," "beautiful" and "ought" behave much like indexical terms. However, a well-known problem for contextualism in these domains is its apparent inability to account for intuitive disagreements.

It has recently become rather fashionable to claim that this problem can be avoided by borrowing an idea advanced by Charles Stevenson and standardly associated with expressivism, viz., that disagreement in certain

1 Thanks to Voin Milevski for inviting me to contribute to this issue. I'd also like to thank everyone who at some point or other provided comments on earlier drafts of this paper. This research was funded by Riksbankens Jubileumsfond (RJ) (grant number: P16-0710:1).

 DOI: 10.5937/BPA1932093E

domains should be understood in terms of the parties having conflicting attitudes. This allows for two parties, A and B, to agree in belief about p, yet disagree in virtue of having conflicting attitudes towards p. In this paper, focus is on moral contextualism and moral disagreement.[2] Given the assumption that moral disagreement is best understood as disagreement in attitude and that the contextualist explanation works, one of the most trenchant objections is thus circumvented.[3] Moreover, it also undermines the view that intuitive disagreement in the absence of disagreement in belief provides one-sided support for expressivism. This paper examines a number of different ways in which disagreement in attitude can be incorporated into a contextualist framework and argues that each of these ways lead to problems. In particular, the purported explanations of disagreement fail to adequately explain intuitive occurrences of disagreement, the robustness of disagreement intuitions and/or locate the disagreement in the intuitively right place.

The outline of this paper is as follows. In the next section the standard objection to contextualism is explained. Section 2 introduces the distinction between disagreement in belief and disagreement in attitude and explains how expressivists make use of the latter. Section 3 introduces the basic contextualist maneuver aiming to accommodate disagreement intuitions in terms of disagreement in attitude. In section 4 through 7 I undertake more detailed examinations of contextualist explanations of disagreement in terms of disagreement in attitude but argue that they all fail.

1. Contextualism and disagreement

Contextualism is the view that the meaning, reference or truth conditions of a class of sentences depend on features of the context, e.g., place, time or the standard of the judge. For example, the meaning of a sentence involving "here" depends on the place of the speaker. Consider the following short exchange involving John and Jane.

(1) It's hot here.
(2) It's not hot here.

Unless we know that John and Jane are in two different places, we may intuit them as disagreeing (at least assuming that they use the same standard for

2 Although focus is on moral contextualism and moral disagreement, the considerations advanced in this paper will most likely also generalize to contextualism in other domains where contextualism has similar problems with respect to disagreement and where similar solutions are proposed.

3 This paper will simply assume that disagreements in the relevant domains are plausibly thought of as disagreements in attitude (rather than disagreement in belief). Moreover, it also assumes that e.g., approval and disapproval of the same subject are states of mind that disagree. Without such assumptions, appealing to disagreement in attitude would be a non-starter for the contextualist.

"hotness"). However, if we learn that John is in Los Angeles while Jane is in Alaska, what (1) and (2) really mean is roughly the following.

(1*) It's hot in Los Angeles.

(2*) It's not hot in Alaska.

Given that (1) and (2) are uttered in two different places, there doesn't seem to be any sense in which John and Jane disagree (as (1*) and (2*) should make evident). In other words, any sense of conflict should disappear. Moreover, there is nothing odd about this. Consider instead the following example where Mary and Mark ponder whether Huck ought to tell on Jim (the fugitive slave) or not and come to the following conclusions.

(3) Huck ought to tell on Jim.

(4) Huck ought not to tell on Jim.

It seems that Mary and Mark disagree.[4] Given an invariantist outlook, for example, two beliefs are in conflict if they (or their content) cannot be true simultaneously. However, if contextualism is correct, then it seems that there is no explanation of the conflict. Rather, we arrive at the following rough semantic interpretations.

(3*) Huck ought to tell on Jim relative to Mary's moral standard.

(4*) Huck ought to not tell on Jim relative to Mark's moral standard.

Given these interpretations, there doesn't seem to be any conflict between Mary and Mark's beliefs. For example, it is true that Huck ought to tell on Jim relative to Mary's moral standard and simultaneously true that Huck ought not to tell on Jim relative to Mark's moral standard. Hence, the disagreement is lost. However, by contrast to the indexical example above, the sense of disagreement doesn't go away. Insofar as we ascribe to Mary and Mark the relevant moral beliefs, we seem to think that they disagree, i.e., we intuit that there is a conflict between the parties' respective views. The challenge for contextualists is to find some way of making sense of this.[5]

2. Disagreement and attitudes

Issues regarding disagreement play an important role in many areas of philosophy. Famously, moral expressivists have argued that their analyses gain

4 It should be emphasized that "disagreement" is a term that can be used in many different ways. It can be used to say that two parties simply have different views, but the sense relevant here is that their views are somehow in conflict. This is the datum that needs explanation.

5 It should be noted that the idea of making sense of disagreement by using the standard expressivist story also has been suggested in other domains, e.g., taste. Although one may argue that disagreement intuitions regarding matters of taste are less robust or somewhat different, the problems raised in this paper apply generally to such attempts to explain disagreement intuitions. See e.g., Eriksson (2016) for discussion.

support from the apparent possibility of agreement on all factual matters (or agreement in belief), yet disagreement in moral judgment.[6] Interestingly, one of the key motivations for some kind of contextualism is also a key part of the expressivist argument from disagreement. One of the premises in the standard expressivist argument is that moral terms like "right," "wrong" or "ought" are subject to systematic variation. For example, we may recognize that Mary and Mark use the terms to systematically pick out different properties. This suggests that the terms have different descriptive meanings in their respective idiolects, but we don't think that this is something that makes either Mary or Mark linguistically confused.[7] However, if the meaning of moral predicates is context sensitive and different in different peoples' idiolects, then it seems as if it will be difficult to explain the intuitive sense of disagreement. The next move made by expressivists is to argue that apparent disagreement in belief isn't the only sense of disagreement. As Stevenson famously pointed out, we must distinguish between disagreement in belief and disagreement in attitude. A disagreement in belief regarding a certain question occurs in cases where "one man believes that p is the answer, and another that not-p, or some proposition incompatible with p, is the answer" (Stevenson 1944: 2). For example, if I believe that Paris is in France while you believe that Paris is not in France, you and I disagree in belief. A disagreement in attitude, by contrast, is characterized as follows.

> Two men will be said to disagree in attitude when they have opposed attitudes to the same object—one approving of it, for instance, and the other disapproving of it—and when at least one of them has a motive for altering or calling into question the attitude of the other. (Stevenson 1944: 3)

Note that this characterization seems to involve two different conditions, but one may think that the motive for altering or calling into question the attitude of the other isn't strictly speaking necessary. Indeed, in other passages, this condition is omitted. Consider instead the following passage that also addresses the difference between the two senses of disagreement.

> The difference between the two senses of "disagreement" is essentially this: the first involves an opposition of beliefs, both of which cannot be true, and the second involves an opposition of attitudes, both of which cannot be satisfied. (Stevenson 1962: 2)[8]

6 This is a prominent argument amongst philosophers in the expressivist tradition. See e.g., Stevenson (1944, 1963), Hare (1952), Gibbard (1990 ch.1), Blackburn (1984: 168, 1991) and Horgan and Timmons (1991). See also Tersman (2006) for a general discussion of moral disagreement. Ayer (1936: 110), by contrast, denies that we disagree about values.

7 Confer Tersman's latitude idea (see Tersman 2006).

8 It is the latter conception that has been the most influential in the development of expressivism (see Ridge 2013: 44). See also Eriksson (2016) for discussion.

Distinguishing between disagreement in belief and disagreement in attitude opens up conceptual space. Most importantly, it makes it possible to agree in belief, yet disagree in attitude. For example, given that Mary and Mark use "ought" in systematically different ways it seems plausible to think that they will not disagree in belief. It is, in other words, conceivable that they agree on all factual matters. Hence, we need some other way of making sense of the disagreement. Of course, it is at this point that disagreement in attitude becomes important. Although Mary and Mark may agree about all factual matters, it is nevertheless possible that they have opposed attitudes towards telling on Jim. Mary approves of telling on Jim. Mark, by contrast, disapproves of telling on Jim. Consequently, Mary and Mark disagree in attitude. This is what, according to the expressivist, explains the disagreement (i.e., appearance of conflict). Moreover, since a moral disagreement isn't a disagreement in belief, we also have reason to think that moral beliefs aren't beliefs with a mind-to-world direction of fit, but noncognitive states. To believe that one ought to tell on Jim is (roughly) to approve of telling on Jim. To believe that one ought not to tell on Jim is (roughly) to disapprove of telling on Jim.

3. Contextualism and disagreement in attitude

Intuitively, Mary and Mark disagree. This is a kind of datum that needs to be explained. The problem for contextualism is that it seems that Mary and Mark aren't really disagreeing since their respective moral views are consistent. However, given that the moral domain also seems to be intimately associated with attitudes and attitudinal expression, perhaps the relevant kind of disagreement is best understood in terms of disagreement in attitude. Although this is an argument that is most intimately associated with expressivism, perhaps contextualists simply can use the same explanation.

The basic idea begins by arguing that the challenge to contextualism rests on a too narrow conception of disagreement. Contextualism, common lore tells us, is unable to account for disagreement. However, this claim has bite only if it is assumed that contextualists must explain disagreement in terms of conflicting propositions or disagreement in belief, but this isn't the only sense of disagreement available. Rather, even if there is no disagreement in belief between Mary and Mark, even if they don't accept inconsistent propositions, there may nevertheless be a disagreement in attitude.[9] In fact, this is an idea that is widely endorsed. Let me run through some examples.

9 This is not the only way in which one may try to avoid the standard objection to contextualism. For example, one may argue that although the propositions aren't literally inconsistent, there is nevertheless a proposition that is picked out as contextually salient that they disagree about. Consider an example: Suppose that I judge that A has one child. You judge that A has two children. It may be argued that you and I don't express logically

Supposing that the expressivist manages to account for disagreement in terms of disagreement in attitude, then, as James Dreier, claims "the indexical theorist may say just the same thing that the expressivist says, namely, that there is real disagreement in norms, or in attitude" (Dreier 1999: 569) and "the account of conflict of attitudes can be adopted by Indexical Relativism" (Dreier 2009: 107). Teresa Marques claims that contextualists should not account for disagreement in doxastic terms, but "turn to the incompatibility of non-doxastic attitudes" and that "[t]he existence of non-doxastic disagreement is compatible with a standard form of contextualism" (Marques 2014: 140). Timothy Sundell, similarly, claims that inconsistent propositions are quite irrelevant because "the conflicting attitudes that the speakers express is all that is required to explain their 'taking themselves to disagree'" (Sundell 2011: 282). More generally, but in the same spirit, Frank Jackson and Philip Pettit make the following claim.

> Indeed, almost every party to the debate in meta-ethics believes that if I sincerely assert that X is right and you sincerely assert that X is wrong, we must have different moral attitudes; so, if that counts as our disagreeing, as expressivists who are not eliminativists about moral disagreement must allow, almost every party to the meta-ethical debate can respond to the problem of moral disagreement simply by noting that a difference in moral attitudes can survive agreement over all the facts. (Jackson and Pettit 1998: 251; see also Jackson 2008)

Other philosophers who have suggested similar ideas include David Wong (1986), Gilbert Harman (1996: 33–37), Gunnar Björnsson and Stephen Finlay (2010), Finlay (2014a, 2014b), Andy Egan (2010), Torfinn Huvenes (2012),

---

inconsistent propositions because my judgment expresses the proposition that A has at least one child while you express the proposition that A has at least two children. In this case, both you and I are right, because A has two children. The moral of such examples is that disagreement in talk doesn't require that the speakers are expressing mutually inconsistent propositions (Björnsson and Finlay (2010), Sundell (2011), Plunkett & Sundell & (2013), Egan (2014)). This seems right, but also quite trivial. We don't always intuit a disagreement in virtue of the literal proposition expressed, but in virtue of what is intended to be communicated. It may also be argued that the salient standard picked out isn't necessarily the speaker's own, but a group standard of some kind (see e.g., Recanati (2007: ch. 11). The example involving height may also be used to illustrate a different way in which a contextualist may want to maneuver around the problem. What is it that we communicate by disagreeing about the tallness of a certain person? A suggestion is that we disagree "about how to use a certain word appropriately" (Barker 2002: 1–2) (see Plunkett and Sundell (2013) for extended discussion on this matter). Lopez de Sa (2009) argues that "is funny" triggers a presupposition of commonality, i.e., roughly that we are similar with respect to humor. Khoo and Knobe (2016) "locates the disagreement between two speakers in their making incompatible proposals to change some aspect of their conversational context" (Khoo and Knobe 2016: 2). I will not here address these suggestions. Rather, the concern is exclusively with trying to account for disagreements in terms of standard expressivist story, viz., in terms of conflicting attitudes.

Teresa Marques and Manuel Garcia-Carpintero (2014). This brief summary is probably far from complete, but it shows how widely endorsed the main idea is.

By accounting for disagreement in terms of disagreement in attitude, one of the most trenchant objections to contextualism is dispelled. Moreover, it also undermines the expressivists' claim that disagreement in the absence of disagreement in belief provides one-sided support for expressivism. As Huvenes claims, "thinking about disagreement in this way doesn't force us to adopt a particular semantic theory. One can think about disagreement in this way without endorsing expressivism" (Huvenes 2012: 179). Arguments from disagreement therefore seem to lack semantic significance. However, the idea that contextualism can make sense of disagreement via disagreement in attitude has not been explored in much detail.[10] Indeed, once one starts examining the idea more closely, it soon becomes clear that it is easier said than done.

4. Disagreement in expressed attitude

According to expressivism, moral assertions function to express rather than report attitudes. This is supposed to be something that shows expressivism to be superior to subjectivism because it promises to explain disagreement intuitions. However, there seems to be no good reason to reject the idea that an assertion can function to both report and express an attitude, e.g., if the latter is expressed pragmatically. Björnsson and Finlay (2010), for example, argue that "ought claims relativized to the speaker's own standard will have the *conversational role* of prescriptions or imperatives" (Björnsson and Finlay 2010: 32; emphasis added). Thus, asserting that Huck ought not to tell on Jim functions to express a proposition (in virtue of its semantics) and a prescription not to tell on Jim (in virtue of the pragmatics). The latter is expressed by virtue of its conversational role. Similarly, Sundell (2011) claims that expressing inconsistent propositions is irrelevant because "the conflicting attitudes that the speakers express is all that is required to explain their 'taking themselves to disagree'" (Sundell 2011: 282). Thus, the basic idea is that two parties disagree because their assertions (in part) function to express conflicting attitudes.[11] Call this disagreement in expressed attitude:

> Disagreement in expressed attitude: A's assertion that p disagrees with B's assertion that q in virtue of A and B's respective assertions expressing conflicting attitudes.

At first glance, this may seem as a plausible way of explaining disagreement intuitions for contextualists. However, there are also some questions that needs to be addressed in order to more fully assess the proposal. One question

10 Köhler (2012) is a notable exception.

11 C.f., Finlay (2014a: 134).

concerns the expression relation: how does an assertion express an attitude? Björnsson and Finlay seem to think of the expression relation as roughly similar to how conversational implicatures work. Others may think that the expression relation is more a matter of convention.[12] However, regardless of which option one favors, the proposal runs into problems. First, one can express an attitude that one doesn't have. Second, one can express an attitude that one thinks one has, but be mistaken about this. These possibilities raise questions regarding whether disagreement in expressed attitude actually explains intuitive occurrences of disagreement in the right way.

The problem with disagreement in expressed attitude is, on the one hand, that we risk failing to explain disagreement where they intuitively occur and, on the other hand, that we find disagreement where there intuitively are none. Of course, there will be cases in which we may disagree whether two parties disagree or not. The following two cases, however, are hopefully cases where disagreement intuitions are uniform.

First, suppose that Mark believes that one ought not to tell on Jim and Mary believes that one ought to tell on Jim. Mary, however, is self-deceived and falsely believes that she believes that one ought not to tell on Jim. She therefore asserts that one ought not to tell on Jim. Of course, this is also what Mark assert. Consequently, in virtue of the parties' respective assertions, both express disapproval of telling on Jim. Given disagreement in expressed attitude, it thus seems that there is no disagreement. However, given that we know that they actually have the beliefs that they have, this seems wrong. Mary and Mark intuitively disagree. Second, suppose that Mary doesn't really think that one ought to tell on Jim, but that she merely wants to examine the issue.[13] Mark, however, is not aware of this. Hence, in virtue of their assertions, Mary and Mark express approval and disapproval of telling on Jim respectively. If disagreement intuitions are supposed to be explained in terms of expressed attitudes, it seems that Mary and Mark disagree. However, given what we know about their respective beliefs, this seems wrong. Hence, the contextualist appeal to disagreement in expressed attitudes doesn't seem to work. It fails to adequately explain intuitive occurrences of disagreement in the right way.[14]

---

12 See e.g., Copp (2001, 2009) for arguments along these lines. One may also think that this brings about a too tight connection to attitudes (see e.g., Finlay 2004)

13 Suppose, for example, that Mary thinks that the only way of getting Mark to engage seriously with the question is by understanding her speech act as an assertion and that she thus manages to assert, albeit insincerely, that one ought not to tell on Jim. If you are inclined to think that Mary, even given her motive, fails to express the relevant attitude, simply assume that she mistakenly thinks that she has the belief in question.

14 It may be argued that the argument against the success of the contextualist explanation of disagreement intuitions in this paragraph trades on an ambiguity, i.e., whether the aim is to explain the intuitive disagreement between, on the one hand, the two parties or, on the other hand, between the assertions they make. Although a complete theory about

In order to help diagnose the problems raised above, we should distinguish between at least two possible senses of disagreement, viz., disagreement in thought and disagreement in talk.[15] On the one hand, there is a sense in which we intuit a conflict between two parties whose assertions express conflicting attitudes. On the other hand, there is a sense in which we intuit a conflict between two parties who simply have conflicting attitudes. Contextualists tend to focus on the former: they tend to be concerned with explaining disagreement in talk. However, as the two cases above illustrate, intuitions regarding disagreement in talk and disagreement in thought can come apart. Two persons can express conflicting attitudes despite not *having* conflicting attitudes (and vice versa). Moreover, even if both senses of disagreement deserve to be called senses of disagreement, the examples above also suggest that disagreement in thought is more fundamental.[16] It is perhaps plausible to think that the parties will intuit that they disagree in virtue of the expressed attitudes. However, it also seems plausible to think that they will, upon discovering that one party was self-deceived or for some other reason doesn't have the attitude he or she expresses, stop thinking of themselves as disagreeing. In other words, upon realizing that they don't have conflicting attitudes, they don't discover something that resolves their disagreement. Rather, what they discover is that they never really disagreed to begin with – although they thought they did because they thought that their respective assertions reflected their actual views. As Jackson and Pettit claims "the production of moral sentences makes public our disagreement; it does not create them" (Jackson and Pettit 1998: 251).

5. Disagreement in attitude

Disagreement in expressed attitudes went wrong because of its focus on the attitudes expressed rather than the attitudes actually had by the two parties. The latter idea is more in line with the standard Stevensonian or expressivist conception of disagreement: two parties disagree if they actually *have* conflicting attitudes. However, this idea also seems possible to combine

---

disagreement should explain both, it is the former that I have in mind and that I take to be a problem for the contextualist. The point is simply that judgments about disagreement in thought and talk (more on this distinction below) can come apart and that the latter therefore cannot fully explain the former. Of course, if the contextualist merely wanted to explain disagreement in assertion, then these objections can be disregarded. However, if this is the aim, I don't think much have been done to solve the disagreement problem for contextualism.

15 See, e.g., Egan (2014: 76). Similarly, Cappelen and Hawthorne (2009: 60–61) distinguish between disagreement as a state and disagreement as an activity. The latter sense requires that two parties are having a disagreement, i.e., that they are in some sense interacting, e.g., in an argument, discussion or the like. The former sense, by contrast, doesn't require that the disagreeing parties interact. Rather, it suffices that the parties have conflict beliefs.

16 See also MacFarlane (2014: 119–120) and Marques (2014) for similar views.

with a contextualist theory. Consider, for example, the following idea pursed by Huvenes.[17]

> The idea I am interested in is to view disagreement as a matter of the parties' *having* incompatible or conflicting attitudes. Two parties disagree just in case there is something towards which they *have* conflicting attitudes. (Huvenes 2012: 178–179; my emphasis)

Call this view Actual attitudinal disagreement.

> Actual attitudinal disagreement: A disagrees with B about p in virtue of A and B *having* conflicting attitudes towards p.

In fact, according to more traditional forms of contextualism, there is an intimate link between moral beliefs and attitudes. According to simple speaker relativism (aka subjectivism), to believe that one ought not to tell on Jim is to believe that one disapproves of not telling on Jim. One way of interpreting Dreier's contextualism (or indexical relativism as he calls it) is that the meaning of moral predicates is determined by the judge's moral standard where a moral standard is identical with (a set of) motivational attitudes. For example, to have a utilitarian standard is, roughly, to approve of maximizing happiness. Given such a moral standard, "Donating to charity is right," in the judge's idiolect, means that donating to charity maximizes happiness. These forms of contextualism explain why there is indeed a very close connection between moral beliefs and attitudes. This, in turn, may very well be one explanation of why it, to many parties, seems easy for a contextualist to make use of opposed attitudes to explain disagreement.

However, given a contextualist theory, it nevertheless seems conceivable to believe that Huck ought to tell on Jim without any concomitant approval of telling on Jim. As Huvenes writes, "a sincere utterance of [one ought to tell on Jim] is typically, *though not invariably*, accompanied by the speaker's having a certain attitude towards [telling on Jim]" (Huvenes 2011: 179 emphasis added). In other words, approval of telling on Jim is contingent. Consequently, if either (or both) party(ies) lacks the required attitude, the explanation of why they disagree disappears.

Nevertheless, given that Mary and Mark have their respective beliefs, there is still an appearance of disagreement. This takes us back to a point made in the beginning of this paper. By contrast, to the indexical example involving "It is hot here," where the appearance of disagreement disappears once we learn that the speakers are in different places, moral disagreement intuitions are much more robust. As long as the two parties' relevant moral beliefs are in place, we intuit that they disagree.

---

17 Huvenes focuses on predicates of taste, but I take it that the general idea can also be used in relation to moral predicates.

It may be claimed that the moral disagreement intuitions aren't as robust as I claim them to be. Let me, therefore, quickly point to two different considerations that suggest that they are. First, if I discover that you are in a different place than I am, it clearly is infelicitous to signal disagreement by saying "No, it isn't hot here" or the like. By contrast, it always seems felicitous to signal disagreement with, e.g., Mary's moral judgment by saying "No, one ought not to tell on Jim" or the like.[18] Second, there is empirical data that supports the modal robustness intuition. Justin Khoo and Joshua Knobe (2016), for example, advance considerations that purport to show that we don't necessarily intuit that at least one party of a moral dispute is wrong or mistaken.[19] However, they nevertheless find evidence for thinking that disagreement intuitions don't go away. These two considerations suggest that moral disagreement intuitions are robust. Hence, in so far as we attribute Mary and Mark with the moral beliefs we have been toying with, it seems that people intuit that they disagree. The challenge is to make sense of this.[20]

One may think that this challenge is easily met by speaker relativism and/or the view attributed to Dreier above. This, however, isn't the case. According to speaker relativism, to believe that one ought to tell on Jim is to believe that one approves of telling on Jim. Although the accessibility to our own minds may perhaps be privileged, we are not infallible. Mary may believe falsely that she approves of telling on Jim whereas Mark believes (correctly) that he disapproves of telling on Jim, which is tantamount to Mark believing that one ought not to tell on Jim. Given the parties' respective beliefs, i.e.,

---

18 I may be argued that the use of disagreement markers merely matters to disagreement in talk. However, it seems that the use of disagreement markers is a way of signaling disagreement with the person. If Mary asserts that one ought to tell on Jim and Mark responds "No, one ought not to tell on Jim," then it seems plausible to think that Mark disagrees with Mary (or her belief) and not merely with her assertion. Rather, Mark takes Mary's assertion to be indicative of her belief on the matter, which is what he really disagrees with. Although the relation between disagreement in thought and talk is in need of a more thorough examination, it clearly seems that disagreement markers isn't merely relevant to disagreement in talk. Rather, most of the time, as Jackson and Pettit claim, disagreement in talk makes public disagreement in thought.

19 This constitutes an interesting challenge to both realists and quasi-realists, but this is an issue that we can set aside for the purpose of the present paper.

20 One may think that Khoo and Knobe's results are irrelevant in the present contexts since they haven't tested whether people's disagreement intuitions would be affected if we were to stipulate that the parties lacked the relevant attitudes. This is, of course, true. However, I very much doubt that people's intuitions would be affected by such a stipulation. On the one hand, insofar as we ascribe to, e.g., Mary and Mark, the moral beliefs we have been toying with, I predict that most people will intuit that they disagree. On the other hand, stipulating that the parties lack the relevant attitudes, may lead people to intuit that they don't disagree. However, I also hypothesize that this will be because this will interfere with ascribing to the parties the moral beliefs to begin with. Nevertheless, the main point about the Khoo and Knobe argument is that disagreement intuitions seem quite robust, but I grant that this intuition may be proved wrong.

that Mary believes that one ought to tell on Jim and that Mark believes that one ought not to tell on Jim, we intuit that they disagree. However, since the parties don't have conflicting attitudes, the explanation of disagreement in terms of conflicting attitudes doesn't work. Again, the disagreement is lost.

A slightly different problem arises for the kind of view Dreier advances. Begin by considering how the content of a moral term is determined.

> The content of a moral term is a function of the affective attitude of the speaker in the context. Thus, "x is good" means "x is highly evaluated by standards of system M," where M is filled in by looking at the affective or motivational states of the speaker and constructing from them a practical system. (Dreier 1990: 9)

Given that the meaning of "good" is a function of the affective attitude it may seem as if we will end up with a view according to which someone who believes that stealing is wrong will necessarily disagree in attitude with someone who believes that stealing is right. This, however, is not the case. Suppose Allan believes that stealing is wrong whereas Brenda believes that stealing is right.[21] Suppose Allan's moral standard is a utilitarian one, i.e., he disapproves of not maximizing happiness. Brenda, by contrast, is of a more Kantian bent and approves of promoting autonomy. These two standards determine the meaning of "right" and "wrong" in their respective idiolects. Allan believes that stealing doesn't maximize happiness. Brenda believes that stealing promotes autonomy. These beliefs, of course, don't disagree. However, neither is disapproval of not maximizing happiness opposed to approval of promoting autonomy. Rather, in order for Allan and Brenda to disagree in attitude they must acquire more particularized attitudes, viz., disapproval of stealing and approval of stealing respectively. The problem is that the acquisition of these attitudes seems contingent. For example, either (or both) party(ies) may fail to acquire the particularized attitude due to some kind of irrationality or psychological failure, but without these attitudes we cannot explain the disagreement as a disagreement in (actual) attitude.[22] Nevertheless, given that Allan believes that stealing is wrong and that Brenda believes that stealing is right, they intuitively disagree. However, since the parties don't have opposed attitudes, the intuition isn't accounted for. Disagreement is, again, lost.[23]

---

21 I will here omit certain complexities of Dreier's view, e.g., that the sometimes is filled in by looking at the motivational states of the larger community. We will return to this issue in section 7.

22 See Eriksson 2015 for discussion on this matter in relation to certain forms of hybrid expressivist theories.

23 Similar considerations seem to be relevant to Finlay's definition "of fundamental disagreements as involving a basic conflict in preferred ends" (2014b: 234). Moreover, it is also not obvious that Dreier escapes the problem addressed above, i.e., if a person is mistaken about his or her standard, then it is conceivable that he or she will fail to have the corresponding attitude (the reason is because one forms one's moral judgment on basis of what one believes about one's moral standard).

6. Disagreement and practical commitments

If the arguments above are right, contextualists run into problems regardless of whether they try to account for disagreement intuitions in terms of expressing conflicting attitudes or having conflicting attitudes since both accounts will fail to explain intuitive occurrences of disagreement. However, maybe the contextualist doesn't have to make sense of disagreement intuitions in terms of either having or expressing conflicting attitudes. Consider the following suggestion:

> Even if strictly speaking our beliefs don't conflict with Huck's, in combination with subscription to conflicting standards these beliefs place us in conflict over the practical matter of what to do in situations like Huck's. In virtue of his subscription to standard Y, Huck's moral belief commits him to favor telling on fugitive slaves. In virtue of our subscription to standard Z, our moral belief commits us to oppose telling on fugitive slaves. Hence these noncontradictory moral beliefs precipitate a disagreement in attitude toward Huck's action. (Björnsson and Finlay 2010: 28)

This idea differs from the previous one advanced by Björnsson and Finlay. First, one can have the particular commitment to an attitude without giving voice to it. Hence, it differs from Disagreement in expressed attitude. Second, the idea doesn't require that the parties have the pertinent attitudes. Hence, it differs from the Disagreement in actual attitude. Rather, the idea seems to be the intuitive disagreement between two parties comes about via their respective commitments to (conflicting) attitudes.[24] Given that the acquisition of the relevant attitudes is contingent, maybe this provides a solution the previous problems for contextualists. Call this conception Disagreement in attitudinal commitment.

> Disagreement in attitudinal commitment: A and B disagree about p if their beliefs together with their moral standard commit them to opposed prescriptions or attitudes (regarding p).[25]

However, since this conception differs significantly the previous two suggestions, it also raises new questions. In particular, we need to examine

---

24 Confer Horwich (2010): "The conflict associated with contradictory beliefs consists in their *potential*, through inference, to engender conflicting desires and decisions. If I disagree with you about the truth of some empirical proposition, <T>, then that can easily result (via theoretical reasoning and given other premises) in our disagreeing about the truth of some more directly action-guiding belief, <If A is done then X will occur>. And if we both want X to occur then one of us will, on that account, be in favor of A being done, and the other won't" (Horwich 2010: 183).

25 Note also that this suggestion differs significantly from more standard conceptions of disagreement in attitude.

how attitudinal commitments come about. We then need to examine whether it handles the problems with the previous accounts.

In order to examine how an attitudinal commitment comes about we must first ask what it is to endorse a certain standard. Björnsson and Finlay's idea seems to be the following: To endorse a standard is to have a preference for some end. To have a utilitarian standard, for example, is to have a preference that happiness is maximized. It is such a preference that fixes Huck's standard. Hence, for Mary to believe that one ought to tell on Jim is for Mary to believe that telling on Jim maximizes happiness. Although Björnsson and Finlay are not entirely clear on exactly how the attitudinal commitment comes about, it seems that the matter is one of simple instrumental rationality.

1. Mary believes that that one ought to tell on Jim
2. Mary has a preference for maximizing happiness.
3. Mary believes that telling on Jim maximizes happiness.
4. Based on 2 and 3, Mary is committed to favor (having a preference for) telling on Jim.

The idea is that Mary is committed to favor telling on Jim because telling on Jim is a means to her end – maximization of happiness – and it is irrational not to favor suitable means to one's ends. Mark's standard, by contrast, is fixed by a preference for some other end, e.g., a preference that agents are to be treated as ends rather than means. What Mark believes when he believes that one ought not to tell on Jim is thus that telling on Jim would be treating him as a means rather than as an end. Hence, the idea is that Mark is committed to having a preference in favor of not telling on Jim. Consequently, Mary and Mark are committed to conflicting attitudes.

Although this proposal is interesting, it also has problems. Plausibly, Mary's believing that telling on Jim is a means to maximizing happiness (i.e., her end), rationally commits her to telling on Jim, not to favor telling on Jim. This is most easily seen by considering the following possibility: although Mary believes that telling on Jim is a means to maximizing happiness, she may simultaneously believe that favoring (having a preference for) telling on Jim will not maximize happiness (because of the consequences of such an attitude). Hence, we can add 3* to 1–4 above:

1. Mary believes that one ought to tell on Jim.
2. Mary has a preference for maximizing happiness.
3. Mary believes that telling on Jim maximize happiness.

3*. Mary believes that having a preference for telling on Jim will not maximize happiness.

4*. From 2 and 3*, Mary is committed to not have a preference for telling on Jim.

In this scenario, it seems that 4* is what we should conclude. In other words, it is rational for Mary to do that which is a means to her end, i.e., to tell on Jim and not to prefer to tell on Jim. Here we can toy with two variations. In one scenario, Mary is committed to no attitude in particular and in a second scenario she is committed to having a preference for not telling on Jim (depending on what we assume that she believes maximizes happiness). Regardless of which route we take, the problem of making sense of intuitive occurrences of disagreement will resurface. Despite thinking that one ought to tell on Jim, Mary is, because of her belief regarding the consequences of favoring of telling on Jim, committed to favoring not telling on Jim (in the latter scenario). This is the kind of attitude that Mark, who believes that one ought not to tell on Jim, is also committed to. Hence, despite the intuitive disagreement between the parties' respective moral beliefs, they are committed to the same attitude. Consequently, the account fails to explain disagreements where they intuitively occur (it will, for similar reasons, also fail to explain agreement where they intuitively occur). It may be objected that I have misunderstood how the attitudinal commitment comes about. Maybe this is true, but it nevertheless remains unclear whether there is some way of explicating this idea that avoids the problems raised above.[26]

7. Disagreement and de dicto internalism

The problem with the suggestions above is that they fail to explain how the moral belief, e.g., that one ought to tell on Jim, necessarily co-occurs in the right way with the right attitudes, i.e., approval of telling on Jim, which is supposed to account for the appearance of disagreement (in terms of disagreement in attitude). A way to try to get around this problem is to consider a move made by Jon Tresan (2006, 2009). Tresan has in different places argued in favor of de dicto internalism with a communal twist. According to this view, a moral belief may be a prosaically factual belief, but in order for the belief to count as a *moral* belief, it must be accompanied by the relevant pro– or conattitudes. Moreover, Tresan argues that moral beliefs need not be accompanied by attitudes at the individual level, but merely at the communal level.[27] In fact, Dreier seems sympathetic to this communal feature. Sometimes the standard isn't filled in by the speaker's motivational attitudes, but "constructed from the attitudes of the larger community" (Dreier 1990: 25).

The nice feature, in this context, is that this move gets us a necessary connection between moral beliefs and attitudes. However, given the

26 Many thanks to Ragnar Francén for helping me think about these matters.

27 It should be noted that Tresan doesn't use de dicto internalism to make sense of moral disagreement, but to account for internalist intuitions. Moreover, Tresan is not a contextualist, but an invariantist.

communal twist, we don't end up with the right result. It may be the case that the content of either party's moral belief is determined by attitudes of the community. If this is the case, then there will be (given contextualism) neither a disagreement in belief nor in attitude – since the judge doesn't have the pertinent attitude. An alternative is, of course, to drop the communal twist in favor of a strict individualistic de dicto internalism. On such a view, a moral belief may be a prosaically factual belief, but it counts as a moral belief only in so far as it is accompanied by a corresponding attitude at *the individual level*. For example, Mary's belief that one ought to tell on Jim is a moral belief if and only if Mary favors telling on Jim. Mark's belief that one ought not to tell on Jim, by contrast, counts as a moral belief if and only if Mark favors not telling on Jim.[28] This move would seem to enable the contextualist to explain moral disagreement as a disagreement in attitude since a moral belief is guaranteed to be accompanied by a corresponding attitude. This, it seems, would help explain intuitive occurrences of moral disagreement in the right way and explain the modal robustness intuition. However, considering the de dicto move reveals a more general problem with the contextualist attempt to explain disagreement in term of disagreement in attitude. In order to bring out the problem, we need to note some of the important differences between expressivism and contextualism.

All the ideas considered above purport, in one way or other, to make sense of moral disagreement by taking over a key feature of expressivism, viz., that disagreement is to be understood in terms of conflicting attitudes. However, there are still important differences between the doctrines, two of which needs to be emphasized in the present context. The first difference concerns the nature of moral beliefs. According to expressivism, to believe that Huck ought to tell on Jim is to approve of telling on Jim. The attitude is, in other words, part of the moral belief itself. According to contextualism, by contrast, this is not the case. Rather, to believe that Huck ought to tell on Jim is to have a prosaically factual belief the value of which is contextually determined. The attitudinal part is, in other words, not part of the moral belief itself (this is the case even if one adheres the de dicto idea outlined in the previous paragraph). The second difference concerns the semantics. According to expressivism, we should explain the meaning of a sentence in terms of the attitude it expresses. Hence, the meaning of, e.g., "Huck ought to tell on Jim" should be understood in terms of the state of mind that the sentence functions to express, e.g., approval of Huck telling on Jim. According to contextualism, by contrast, this is not the case. The attitude is not part of the semantics of the sentence. Rather, the meaning of the sentence is exhausted

28 A question in relation to the de dicto idea is also what kind of attitude that a moral belief needs to be accompanied by. For example, suppose Jack is a utilitarian, believes that stealing fails to maximize happiness and therefore believes that stealing is wrong. Does his belief count as a moral belief only it is accompanied by disapproval of stealing or does it suffice that he disapproves of not maximizing happiness?

by the proposition expressed. The attitudinal part is merely pragmatics. It is partly because of these differences that the connection between moral belief and attitude is contingent and risks failing to explain intuitive occurrences of disagreement in the right way. The de dicto idea avoids that problem. In order to see why this presents a problem for contextualist views, let us first consider the disagreement between Mary and Mark again.

(3) Huck ought to tell on Jim.

(4) Huck ought not tell on Jim.

Intuitively, Mary and Mark disagree in virtue of accepting these two respective moral beliefs and these two beliefs alone. In other words, moral disagreement intuitions are not merely modally robust, but the disagreement, i.e., the sense of conflict, is due to *a conflict between the relevant moral beliefs*.

In order to try to bring out this intuition more clearly, consider how we think about prosaically factual disagreement. Suppose John believes that Paris is the capital of France whereas Jane believes that Paris is not the capital of France. John and Jane clearly disagree. Such a disagreement is due to the fact that their respective beliefs cannot be true simultaneously. If there is no conflict between their respective beliefs, there is no disagreement regarding the capital of France. In this case, the disagreement is rather obviously located between the respective beliefs alone. Similarly, it seems intuitive to think that moral disagreements are due to two parties having either moral beliefs that cannot be true or false simultaneously or because they have moral beliefs that are constituted by attitudes that are in conflict. Below I will advance two further considerations in support of this.

First, consider someone whom you disagree with on a moral issue. In virtue of what do you, intuitively, disagree with that person? Presumably, you believe that that person has a moral belief that conflicts with your moral belief – not that that person has some other attitude or belief that conflicts with yours. For example, in order for the disagreement to be resolved, the person you disagree with will have to relinquish the moral belief in question. Second, consider the use of disagreement markers. For example, if Mary asserts that one ought to tell on Jim, I take this to express a moral belief that she endorses. You can express disagreement in a number of different ways, e.g., by saying "That's false," "You are mistaken" or "No, one ought not to tell on Jim." What is it, intuitively, that you think is false, mistaken or that you are somehow challenging by responding in this way? Again, it seems that the target of your disagreement is the moral belief Mary has or gives expression to. This seems to suggest that the disagreement intuitively is located between your respective moral beliefs. All in all, the considerations advanced above suggest that we intuit that two parties disagree because they have *moral beliefs that are in conflict*.

If we want to explain the disagreement intuition in terms of disagreement in attitude, then the contextualist story seems to get things wrong. The reason is that the attitude is not part of the moral belief itself. For example, there is no conflict between Mary and Mark's moral beliefs. As Björnsson and Finlay write, "strictly speaking [their] beliefs don't conflict" (2010: 28). Rather, the disagreement intuition is supposed to be explained by something other than your moral beliefs. Consequently, the contextualist will fail to locate the disagreement in the right place. Again, when two people disagree on a moral issue, the disagreement seems to be due to some feature of their respective moral beliefs, but since the attitudinal part, according to contextualist views, isn't part of the moral belief itself, this intuition isn't accounted for. The de dicto view doesn't avoid this problem. Although it is true, given such a view, that a moral belief is always accompanied by the relevant pro– or con-attitude, the attitude is not part of the moral belief itself, which is what we intuitively disagree with. Consequently, contextualism (in any guise) will fail to make sense of the intuitive location of the disagreement, i.e., that the disagreement is due to the parties having conflicting moral beliefs. Expressivists, by contrast, think that moral beliefs are constituted by the relevant attitudes that are in conflict and thus locate the disagreement in the intuitively right place, i.e., as a disagreement between the two parties' moral beliefs.

The contextualist could, of course, argue that there is some way to explain these intuitions away or provide an error-theory regarding the location intuition.[29] However, until this has been satisfactorily done, we have reason to think that the attitude is part of the moral judgment – assuming that we think that moral disagreement is best accounted for in terms of conflicting attitudes. Moreover, and more generally, this also shows that the contextualist cannot simply take over the standard expressivist explanation. Rather, explaining disagreement intuitions using the disagreement in attitude idea requires much more work if it is to fly within a contextualist framework.

Concluding remarks

A standard objection to moral contextualism is that such a thesis cannot make sense of moral disagreement. This paper has considered a popular suggestion advanced in the literature, viz., that contextualists simply can adopt the standard expressivist story. We should not think of the disagreement between two parties as a disagreement in belief, but as a disagreement in attitude – thus mimicking the expressivist idea that there can be moral disagreements without disagreement in belief. This paper has argued that this is easier said than done. In fact, if the arguments of this paper are right, we have reason to be skeptical about its success. This paper has examined

29 For example, upon being told that the disagreement is due to the necessarily accompanying attitude, I still intuit that our moral beliefs are in conflict. Either this is due to some error on my part or the error is locating the disagreement in the wrong place.

a number of different ways in which disagreement in attitude can be incorporated into a contextualist framework all of which lead to problem: the purported explanations of disagreement fail to adequately explain intuitive occurrences of disagreement, the robustness of disagreement intuitions and/or locate the disagreement in the intuitively right place.

Of course, there are other ways of trying to account for disagreement intuitions (see footnote 8) within a contextualist framework. Moreover, one may think that the standard expressivist account is seriously flawed (Ridge 2013, 2014) and think that there are better alternatives. However, the purpose of this paper is merely to examine contextualism in conjunction with the standard expressivist account, i.e., disagreement in attitude. Examining other alternatives is outside the scope of this paper. Nevertheless, those of us who think that disagreement in attitude is a plausible account of disagreement (even if it requires some tinkering) and who think that certain domains are characterized by being intimately connected to nondoxastic attitudes, e.g., ethics and taste, still have reason to think that expressivism is superior to rival theories. If this is right, arguments from disagreement may still have, at least some, semantic significance and thus push us in the direction of expressivism.

References

Ayer, A. J. (1936). *Language, Truth and Logic*, 2nd Edition, London: Victor Gollancz Ltd.

Barker, C. (2002). "The dynamics of vagueness." *Linguistics and Philosophy* 25(1): 1–36.

Björnsson, G. and Finlay, S. (2010). "Metaethical contextualism defended." *Ethics* 121: 7–36.

Blackburn, S. (1984). *Spreading the Word: Groundings in the Philosophy of Language*, Oxford: Oxford University Press.

Blackburn, S. (1991). "Just Causes." *Philosophical Studies*, 61: 3–17.

Copp, D. (2001). "Realist-expressivism: A neglected option for moral realism," *Social Philosophy and Policy* 18: 1–43.

Copp. D. (2009). "Realist-Expressivism and Conventional Implicature" in *Oxford Studies in Metaethics* Vol. 4, ed. R. Shafer-Landau: 167–202. Oxford: Oxford University Press.

Dreier, J. (1990). "Internalism and Speaker Relativism." *Ethics* 101(1): 6–26.

Dreier, J. (1999). "Transforming Expressivism." *Noûs* 33(4): 558–72.

Dreier, J. (2009). "Relativism (and Expressivism) and the Problem of Disagreement." *Philosophical Perspectives* 23: 79–110.

Egan, A. (2010). "Disputing about taste" in *Disagreement*, Richard Feldman and Ted A. Warfield (eds), Oxford: Oxford University Press.

Egan, A. (2014). "There's something funny about comedy: A case study in faultless Disagreement." *Erkenntnits* 79: 73–100.

Eriksson, J. (2015). "Explaining disagreement: a problem for hybrid expressivists." *Pacific Philosophical Quarterly* 96(1): 39–53.

Eriksson, J. (2016). "Attitudinal complexity and two senses of disagreement in attitude." *Erkenntnis* 81(4): 775–794.

Finlay, S. (2004). "The conversational practicality of value judgments." *The Journal of Ethics* 8: 205–223.

Finlay, S. (2014a). *Confusion of Tongues: A Theory of Normative Language*. Oxford: Oxford University Press.

Finlay, S. (2014b). "The pragmatics of Normative Disagreement," in *Having it both ways*, Guy Fletcher and Michael Ridge (eds), Oxford: Oxford University Press: 124–148.

Gibbard, A. (1990). *Wise Choices, Apt Feelings*. Oxford: Clarendon Press.

Gibbard, A. 2003. *Thinking How to Live*, Cambridge, Mass.: Harvard University Press.

Hare, R. M. (1952). *The Language of Morals*. Oxford: Clarendon Press.

Harman, G. (1996). "Moral Relativism" in *Moral Relativism and Moral Objectivity*, Gilbert Harman and Judith Jarvis Thomson, Cambridge, Mass., Blackwell.

Horgan, T. and M. Timmons (1991). "New–Wave Moral Realism meets Moral Twin Earth" in John Heil (ed.), *Rationality, Morality, and Self–Interest*. Lanman Md: Rowman and Littlefield, pp. 115–33.

Horwich, P. (2010). *Truth-meaning-reality*. Oxford: Clarendon Press.

Huvenes, T. (2012). "Varieties of Disagreement and Predicates of taste." *Australasian Journal of Philosophy* 90(1): 167–181.

Khoo, J. and Knobe, J. (2016). "Moral Disagreement and Moral Semantics" *Noûs* 50(2): 1–35.

Köhler, S. (2012). "Expressivism, subjectivism and moral disagreement," *Thought* 1: 71–78.

López de Sa, D. (2009). "Presuppositions of commonality: An indexical relativist account of disagreement" in Garcia-Carpintero, M. and Kölbel, M. (eds.) *Relative Truth*, Oxford: Oxford University Press.

Jackson, F (2008). "The argument from the persistence of moral disagreement" in *Oxford Studies in Metaethics* 3rd ed, Russ Shafer-Landau (ed.) Oxford: Oxford University Press.

Jackson, F and Pettit, P (1998). "A Problem for Expressivism." *Analysis* 58(4): 239–51.

MacFarlane, J. (2014). *Assessment Sensitivity: Relative Truth and Its Applications*. Oxford: Oxford University Press.

Marques, T. (2014). Doxastic Disagreement." *Erkenntnis* 79: 121–142.

Marques, T. & García-Carpintero, M. (2014). "Disagreement about taste: commonality presuppositions and coordination." *Australasian Journal of Philosophy*.

Plunkett, D and Sundell, T. (2013). "Disagreement and the Semantics of Normative and Evaluative Terms." *Philosophers' Imprint* 13: 1–37.

Recanati, F. (2007). *Perspectival Thought*. Oxford: Oxford University Press.

Ridge, M. (2013). "Disagreement." *Philosophy and Phenomenological Research* 86: 41–63.

Ridge, M. (2014). *Impassioned Belief*. Oxford: Oxford University Press.

Stevenson, C. (1944). *Ethics and Language*. New Haven. Yale University Press.

Stevenson, C. (1962). *Facts and Values*. New Haven. Yale University Press.

Sundell, T. (2011). "Disagreements about taste." *Philosophical Studies* 155: 267–288.

Tersman, F. (2006). *Moral Disagreement*. Cambridge: Cambridge University Press.

Tresan, J. (2006). "De Dicto internalist cognitivism." *Noûs* 40(1): 143–65.

Tresan, J. (2009). "The challenge of communal internalism." *The Journal of Value Inquiry* 43: 179–99

Wong, D. (1984). *Moral Relativity*. Berkeley: University of California Press.

Marko Konjović
Central European University,
Department of Philosophy
University of Belgrade,
Institute for Philosophy and Social Theory
marko.konjovic@instifdt.bg.ac.rs

Original Scientific Paper
UDC 17.022.1:[159.9:17
177.6:17.022.1

REASONS OF LOVE AND MORAL THINKING

Abstract: *There are two widely-held intuitions about morality. One is the claim that all persons have equal moral worth; the other is that sometimes we are morally allowed or even required to give preference to those individuals whom we love. How can we justify our reasons of love in the face of moral egalitarianism? As of recently, there are three mutually competing accounts of why it could be said that we have reasons of love: (i) the projects view, (ii) the relationship view, and (iii) the individuals view. In this paper, I first examine these three views and find fault with each of them as they stand. I then proceed to propose a complex, yet a more compelling, account of reasons of love that builds on the individuals view.*

Key Words: *reasons of love, moral equality, moral thinking, partiality, impartiality*

Introduction

Daily life is infused with making moral decisions. How should we reach them? Consider the following case:

> *Drowning Mother*: You are out on a nice and relaxing Sunday-afternoon stroll alongside the Danube river when you hear a commotion ahead: a splash of water and calls of distress. Running forward, you see two individuals who have fallen into the murky waters. There happens to be one life jacket on the sidewalk, which you grab while running towards the river's bank. Because the river is moving quickly, you will only be able to throw the life jacket to one of the individuals who has fallen in and save her; the other individual, sadly, will be carried away by the river. As you reach the river's bank, you realize, with great shock, that one of the individuals in the river is your beloved mother. Without another thought, you throw the life jacket to your mother and pull her to safety as the other individual – a stranger to you – floats away to her death.[1]

1 This example is a stylized and slightly altered version of the *Drowning Wife* case put forward initially by Charles Fried and made extremely popular by Bernard Williams (1981: 1–19).

 DOI: 10.5937/BPA1932115K

Although you – and most other people – might feel as though you acted rightly, you might nevertheless, upon reflection, worry whether you actually did the right thing. What is your justification for saving your mother rather than the stranger? Call the reasons for being partial towards your mother (your friend, your romantic partner, or your child) *reasons of love*.[2] Love, on any account, demands that we give special consideration to those whom we love; that is, love asks us to give more weight, at least in some circumstances, to the well-being of our loved ones. This is because love involves seeing the beloved in a particular way and having a variety of beliefs about her, including, most notably, beliefs about her specialness.[3] Plausibly, partiality is not only morally permitted, but it is also sometimes morally required, at least in certain contexts.[4] (I say a bit more about this in the concluding section.) What is a matter of dispute, however, is what exactly justifies reasons of love.

The question arises, to a large extent, because a cornerstone of (Western) morality is the moral equality of persons thesis: everyone ought to be treated with "equal concern and respect" (to borrow Ronald Dworkin's (1977) eloquent phrase). The moral equality of persons thesis is not only deeply entrenched into our thinking but it is also germane in the fight against nepotism, sexism and misogyny, homophobia, transphobia, racism, xenophobia, ageism, and ableism (to give quite a few examples). How can we then justify that we are morally permitted or even required to give preference to those whom we love, at least in certain contexts?

The main goal of this paper is to offer the contours of a solution to this puzzle. The justification of reasons of love that I suggest stems from a dissatisfaction with three prominent answers which have been advanced in recent philosophical literature; Simon Keller (2013) helpfully classifies them as: (i) the projects view, (ii) the relationship view, and (iii) the individuals view. In Section 2, I examine these three views and find fault with each of them as they are articulated. I then proceed to offer, in Section 3, an account of reasons of love that builds anew on the individuals view. This story takes inspiration from one of Keller's suggestions regarding the role of relationships in accounting for reasons of love; it gives relationships, however, a different flavor. Moreover, the version of the individuals view I propose goes beyond Keller's inasmuch as it incorporates one somewhat neglected, yet important,

2 Samuel Scheffler (2006) refers to what I call "reasons of love" as "relationship-dependent reasons" while Diane Jeske (2008) calls them "reasons of intimacy." Simon Keller (2013) calls such reasons "reasons of partiality." These differences are, so far as I can tell, only terminological.

3 It is worth noting that I am speaking here about love for concrete persons (so, not about love for co-nationals, for example) and in general (so, regardless of any particular kind of love, such as romantic, friendly, or filial).

4 Nowadays, no one upholds the highly implausible idea that morality requires us to always be impartial, so far as I am aware. Moreover, such extreme impartialism is, in all likelihood, psychologically impossible.

theme of the projects view: the idea of a life worth living. In the concluding segment, Section 4, I briefly consider the issue of when partiality is justified; I suggest that there could be a principled way to distinguish between cases in which reasons of love rightfully reign and in which reasons of love ought to be banished.

Three Love Stories

The justification of reasons of love, on one general and quite popular strategy, need not appeal to any additional moral facts beyond the existence of a loving relationship. There are three such non-reductionist stories behind reasons of love: the projects view, the relationship view, and the individuals view. In this section I discuss in more detail these three proposals and present some of the more serious problems for each. Though the issues each view faces need not be decisive, they do motivate the search for an alternative justification.

Before I proceed, it is important to stress an important commitment which non-reductionists uphold. Namely, they develop their approach as a reaction to the standard reductionist justification of reasons of love: they complain that reductionists fail to offer a justification that is in line with the phenomenology of partiality since they focus on impartial moral principles. That a justification of reasons of love is in accordance with how we experience partiality means that our justification ought to correspond to our motivation (Keller 2013: 25–27). I take this to be a compelling point; however, as I argue in Section 3, there is a plausible account of reasons of love that does not separate our motivation from our justification but which is explicable in terms of impartial moral facts.

The Projects View

The projects view is most prominently defended by Bernard Williams (1981), to whom I owe a great deal for the Drowning Mother example. According to Williams, reasons of love are to be found in facts about me. That is, they refer to the reasons generated by our projects. The argument is, in a nutshell, that insofar as we have reasons to be partial to our own projects, we also have reasons to be partial to our loved ones. To assess the plausibility of the projects view, we first need a clearer idea of what a project is. Williams writes:

> A man may have, for a lot of his life or even just for some part of it, a *ground* project or set of projects which are closely related to his existence and which to a significant degree give a meaning to his life (1981: 12).

Ground projects, or projects for short, on a plausible interpretation, are a set of vital concerns, interests, or goals. These might be outcome oriented (such as finishing a PhD) or ongoing (such as parenting). Ground projects are not, according to Williams, mere desires or preferences because they (i) have a greater influence on our self-understanding, and (ii) they are rooted in a history of commitment.[5] These two differences also explain why we have reasons to prefer our own projects to someone else's: should we abandon our projects, we would lose an important part of ourselves. Morality, however, must not make such a high demand.

The main claim of the projects view is that loving relationships are like our ground projects. Indeed, we typically take our beloved to be a more or less central component of our life and identity: loving someone can come to play an important role in the self-conception of each lover.[6] This applies regardless of whether we think about romantic lovers, friends, or family members. Indeed, a future in which my romantic partner, friend, or mother, to take a few examples, is absent would be very different than my present. So, reasons of love are justified insofar as loving another person takes on the role of a ground project.

The projects view, one might initially object, faces the Focus Objection. There seems to be something disturbing in thinking that I am justified in saving my drowning mother because she is crucial for my ground project of being a good son. That is, it could be thought that ground projects are only self-directed. While some of our ground projects can certainly be only self-directed, this is not the most generous interpretation when thinking about loving relationships. After all, ground projects such as those of being a good son, parent, friend, or lover necessarily involve taking into account the interests of others too. Indeed, I would not be a good son if I only act out of my own interest when I save my mother: I should also consider my mother's interest. Some ground projects, thus, need not be only self-directed; they can be other-directed too.

Nonetheless, the projects view has a pertinent problem. Namely, it is faces the Extension Objection. Let me elaborate. Any account of reasons of love must be able to justify why we are morally permitted, or sometimes even required, to give special treatment to some individuals. Such a story ought to cover various relationships which we generally think are characterized by reasons of love: friendships, romantic relationships, parent-child relationships,

---

5 Indeed, the project of writing a PhD thesis or of parenting is quite different from the desire or preference to have Ben and Jerry's ice cream, for example, as much as one likes ice cream and as much as Ben and Jerry's is a delicious ice cream, especially when compared to other ice creams.

6 Susan Wolf (1992), following Williams, argues that we create and express ourselves as individuals partly through our particular commitments, including our commitments to certain people, and that those commitments are then sources of reasons.

and the relationship between siblings. However, as Keller (2013: 39–41) argues, the projects view is unable to cover those cases in which others do not figure in our ground projects. Consider a realistic relationship between siblings. Imagine you and your brother have very different personalities and as a result you do not play a major role in each other's lives. Nevertheless, you still see each other a few times a year, talk occasionally on the phone, and take a modest but genuine interest in one another. Imagine further that your somewhat estranged brother and a person completely unknown to you are drowning in the Danube. Are you justified in saving your brother instead of a stranger? Most people would say that you are. But, this justification cannot based on the notion of a ground project, for your brother is not a part of any of your ground projects.

Surely, it is possible for an advocate of the projects view to bite the bullet and to admit that the projects view cannot cover all cases. Admitting this theoretical limitation comes at a high price, however, for there are other competing accounts which do not face this challenge. That is, the Extension Objection points out that if the projects view aims to provide a comprehensive justification of our reasons of love, it is problematically incomplete. Moreover, it directs us to look for an explanation elsewhere. Let us now turn to the second attempt to justify reasons of love – the relationship view – which promises to provide a more complete account.

The Relationship View

The projects view, I argued, cannot be extended to cover all cases in which we think we have reason of love. The extension problem which the projects view faces hints at the second plausible way to pin down reasons of love: it points to the ethical significance of relationships themselves. We value our relationships with others, Samuel Scheffler, the most ardent exponent of the relationship view, holds, not (merely) because they help us to achieve some further goals; rather, we (also) value them for their own sake. Indeed, few would deny this insight. To value relationships non-instrumentally, Scheffler continues to argue, means to consider them as the source of our reasons of love (2010: 100–104).[7]

The relationship view could be also interpreted as claiming that some particular facts about the relationship – past, present, or future – provide us with reasons of love. This is, I believe, what Virginia Held has in mind when she writes that the reason why a child honors her father is not because the child thinks that honoring one's father is generally a good thing, but because the particular father is worth honoring for the reasons that can be elucidated by describing the details of the relationship over the years (2006: 79–80). This also appears to be the reading of the relationship view that Diane Jeske

7 See also: Kolodny (2003).

(2008) upholds. Although there can be two readings of the relationship view, a compelling objection affects both interpretations.

The relationship view does not face the extension problem. However, it might be argued that the relationship view, on either interpretation, faces the opposite problem. Call this the Overgeneralization Objection. Namely, it could be thought that the claim is that any relationship will do. Thus, even a person who is in an abusive relationship, it could be objected, could have reasons of love towards her abuser if she values that relationship non-instrumentally. However, if you are in an abusive relationship, it is plausible to hold, you do not have reasons of love towards your wrongdoer; quite the opposite, perhaps, you have reasons to give less weight to her well-being than to the well-being of a stranger. Indeed, few people would hold you morally at fault, I suspect, for not giving more weight to the well-being of your abusive mother in comparison to the well-being of a stranger.[8]

This would not be, however, the most charitable reading of the relationship view. A defender of the relationship view, after all, could deny that we have a reason to value abusive relationships. Indeed, "a relationship that is destructive or abusive," as Scheffler casually mentions towards the end of his paper, "lacks the value that makes it a source of reasons to begin with" (2010: 128). The thought is, then, that the relationships which generate reasons of love are those which are valuable in a particular way or which have valuable aspects. Although we might be tempted to ask in virtue of what a relationship possesses or lacks value or valuable aspects regardless of whether someone values that relationship or not, a defender of the relationship view cannot provide us an answer to this query, for a response would require going beyond the relationship itself. This need not be a knock-down argument against the relationship view, however, because one might be content with leaving the view fairly intuitive.

Even if we accept a rather intuitive version of the relationship view, there is a more pertinent problem with the account. Namely, unlike the projects view, the relationship view faces the Focus Objection. Keller presents a phenomenological version of the focus objection in the following way:

> A person who characteristically thinks of her relationships when she acts well toward others is not someone you would want as a friend or loved one. A friend who is always thinking of improving your friendship, a colleague whose main concern is with the value of collegiality, a parent who thinks mainly of how important it is to have a

8 Cases of such "partiality gone bad" (in lack of a better term) are not only fairly common in real life but they are also philosophically interesting. If we are morally justified in giving more weight to the well-being of our loved ones, are we also morally justified in giving less weight to the well-being of those whom we hate? As most, if not all, philosophers who consider the phenomenon of partiality are focused on cases of favorable treatment of those whom we love (in some sense), I accept this restriction.

> good relationship with his child all of these characters are annoying to have around, and all of them seem to be missing what really matters in their relationships. In a relationship with such a person, you may feel that he cares less for you than for his relationship with you. He cares less for you yourself than for a role that he wants you to fill (2013: 63).

The key point Keller makes, I take it, is that we want our loved ones to act for our own sake. Indeed, my mother would surely be disappointed if she were to learn that I saved her because I feared that our valuable relationship would no longer exist. She would have hoped that I saved her because *she* matters to me. A proponent of the relationship view might be tempted to reply that although we might indeed be motivated by a concern for our loved one, it is the loving relationship that produces the reasons why we are justified in acting in such a way. If she were to make this move, however, an advocate of the relationship view would be separating moral justification (the relationship) from moral motivation (the loved one) (Keller 2013: 63–64). Such a move not only makes the relationships view lose its appeal, it is, more strongly, an unacceptable response on the part of a view which is committed to the idea that the justification of reasons of love should be in accordance with our everyday motivation for being partial to our loved ones.[9]

Since the projects view cannot cover all relevant cases and the relationship view lacks a proper focus, what other justification might we offer for reasons of love? The third account found in the literature – the individuals view – aims to vindicate reasons of love while avoiding these two problems. The following section, hence, is dedicated to the individuals view.

The Individuals View

In her best-known academic book, *The Sovereignty of Good*, Iris Murdoch provides the seed of the third attempt to justify reasons of love: she writes that "love is knowledge of the individual" (1970: 28). Nonetheless, it is Simon Keller (2013) who gives the most elaborate and a highly promising version of the individuals view. The individuals view holds, roughly, that reasons of love "arise from facts about the individuals" (Keller 2013: 79) with whom we share a loving relationship, and not from some relational property.

The basic idea is that persons have certain properties that make them valuable to us. It is that valuable property that gives us reasons of love: they are the appropriate response to the value of individuals. What is that valuable property of individual persons? It cannot be something morally arbitrary, such as beauty, intelligence, or humor. These characteristics may be explanations for liking a person but they cannot be a part of a *moral* justification of reason of love. The best candidates, then, seem to be morally relevant properties like rationality, autonomy, interests, sentience, or capabilities to flourish.

9 Thanks to an anonymous reviewer for pressing me to clarify parts of this section.

Immediately, however, the individuals view seems to face the Indeterminacy Objection. After all, it appears that the account lacks the resources to say that we have reasons of love at all: other individuals, besides my loved ones, have a valuable property like rationality, autonomy, sentience, or capabilities. If this is so, then we have little reason to give preference to any particular individual. Keller is keenly aware of this problem; in order to avoid this objection, thus, he offers a more sophisticated story.

In his elaboration of the individuals view, Keller holds that a person's value is tied up to her having "a particular, distinct perspective on the world" (2013: 142) and then draws on Jonathan Dancy's (2004) idea of 'enablers' to explain why we are justified in giving preference to some particular individuals over others. Enablers affect reasons without being reasons *per se*; that is, enablers are background conditions that explain why something counts as a reason. So, while each person possesses equal value, the fact that I participate in a relationship with my mother puts me in a privileged position to experience, understand, and appreciate her distinctive value as a person. The fact I am in a favorable position to appreciate my mother's value as a person enables her value to provide me with a reason to be partial towards her, just like the fact that you are in a favorable position to appreciate your mother's value as a person provides you with a reason to be partial to your mother. However, since your mother is a stranger to me – I share no relationship with her – I do not have any reason to treat her favorably. Thus, I am morally permitted, if not required, to be partial to my mother and you are morally permitted, if not required, to be partial to your mother on the grounds that participating in a relationship enables their individual value as persons to be known to each of us (Keller 2013: 133–144).

Such a solution to the indeterminacy problem is quite intriguing; yet, Keller concedes that it is ultimately "primitivist" inasmuch as it does not provide any further story as to "*why* the fact that you share a relationship with someone should enable her self-standing value to generate special reasons for you" (2013: 135. Emphasis in the original.). Despite this deficiency, Keller maintains that his view ought to be preferred for it has significant advantages over the other two competing accounts. Indeed, unlike the relationship view, first, the individuals view has a proper focus: the person whom we love. Second, unlike the projects view, the individuals view is comprehensive enough (because it makes use of relationships) to account for a variety of cases in which we typically hold that reasons of love reign, such as the estranged siblings case which presents a problem for the projects view.

While I think that Keller is on the right track, his individuals view is not an overall superior justification of reasons of love. This is because Keller's solution is vulnerable to some important problems. As I argue in the following section, however, there is a version of the individuals view which does not face these issues. Before that, let us see what the problems for Keller are.

By applying the notion of enablers to relationships, first, Keller is driven into rejecting an otherwise plausible principle: namely, he denies that "if two entities have the same kinds of value, then any reasons generated by the value of the first entity must also be generated by the value of the second entity" (Keller 2013: 114). Consequently, Keller has two options. On the one hand, he could deny that the value of individuals with whom I do not stand in a valuable loving relationship does not provide me with reasons for action since the relevant enabling condition is not met. This is clearly not an acceptable route to take as it would amount to a rejection of the idea of moral egalitarianism. Thus, Keller opts for the second path: he accepts that reasons of love – the reasons I have to save my mother – are different in kind from what we can call reasons of justice – the reasons I have to save a stranger (Keller 2013: 114). However, it then becomes unclear how relationships transform one kind of a reason into another and what the relevant difference between the two kinds of reasons is supposed to be. Moreover, it is an unnecessary move: a justification of reasons of love need not lead us to open some difficult questions if we make use of another notion.

Second, if we accept that my reason to save my mother is of a different kind than my reason to save a stranger, reasons of love appear to be morally arbitrary. After all, that I am able to form a relationship with my mother, my friends, or my partner is detemined by morally irrelevant features of both our circumstances in life as well as of people's character. Clearly, I did not choose my mother, nor did my mother choose me. Although in some sense we choose our friends and romantic lovers, this choice is also limited by our morally irrelevant circumstances in life, such as where we live, where we work, what our socio-economic status is, and perhaps even what gender/race/sexual orientation we are, and as well as by amoral properties of persons, such as wit. Indeed, it is difficult to see what morally relevant explanation Keller could offer. As I argue in the following section, however, this issue can be surmounted too if we take a different route.

In accounting for reasons of love, the individuals view is, generally speaking, right to place the focus on individuals. It is also promising because it is comprehensive enough to apply to various cases where reason of love reign. However, Keller's articulation of the individuals view makes reasons of love both mysterious and morally arbitrary. The individuals view can be rendered more plausible if we give up the idea that relationships function as enablers of reasons and uphold the idea that relationships intensify our reasons. Taking this route also hints at a plausible answer as to why it is not morally arbitrary to give preference to those whom we love. The following section is, therefore, dedicated to giving more details about this possible path.

An Alternative Love Story

The guiding question of this paper is: what justifies our reasons of love? That is, what moral reasons do we have to give favorable treatment to those people whom we love? So far, I argued that this question is not adequately answered by appealing to neither (i) the value of ground projects, nor (ii) the value of loving relationship themselves, nor (iii) the enabled value of individuals. Nevertheless, the individuals view presents itself as the most promising strategy. The main taks of this section is to present a different version of the individuals view. While I too begin from the value of individuals to justify reasons of love, I argue that relationships have an intensifying role because of their importance for a life worth living. Though such a version of the individuals view is an impartial account, it is better suited to deal with the problems faced by the version of the individuals view examined in the previous section.

The First Step Towards a Solution: The Value of Individuals and Relationships as Intensifiers

The first step in developing a more satisfactory story about reasons of love is to accept that the value of individuals generates reasons for action. Since each person possesses equal value, we have a reason to treat everyone with equal concern. This is precisely what the moral equality of persons thesis holds. I do not provide an argument for this thesis, as it is widely accepted as an axiom from which contemporary moral and political philosophy must start (Anderson 1999; Dworkin 2000, 2011; Kymlicka 2002; Christiano 2007; and others).[10] But, how do we go about justifying that we are sometimes permitted, if not required, to treat some particular others favorably?

I believe that Keller's idea of using relationships as Dancy-style enablers is a step in the right direction. However, instead of thinking of valuable loving relationships as enablers, it is better to think of them as 'intensifiers.' (It is interesting to note that Keller mentions intensifiers in a parentheses but does not make any use of this idea.) Intensifiers, as the name suggests, increase the strength of already existing considerations that speak in favor of performing an action (Dancy 2004: 41–42). To appreciate the distinction between reasons, enablers, and intensifiers, consider the following two examples.

Imagine that a famous artist is having, for the first time in your lifetime, an exhibition in your home city. That the show is in your home city coupled with the fact that you can afford the ticket enables you to go to see the exhibition. Suppose, however, that tomorrow is the last day of the exhibition. This is not in itself a reason to attend the exhibition: perhaps you do not like

10 Nevertheless, providing a defense of this claim may indeed turn out to be "one of the most profound problems of moral philosophy" (Christiano 2007: 54).

art at all or you simply do not like that particular artist. But, if you do like art or you wish to see the work of that particular artist, then the fact that tomorrow is the last day of the exhibition gives more weight to your reason to attend it.

Or consider a different example. You are shopping for a new shirt for work. As you look around your preferred store, you pick up two shirts. Both fit you well and both are of the same quality and price. You have a reason to buy either one of them. However, one of the shirts is in your favorite color – black – and the other is a color you don't like – red. The fact that one of the shirts is black tips the scale in favor of buying that shirt and not the other one. Still, the fact that the shirt is black is not a reason on its own to buy it: perhaps the shirt does not fit your body type.

The suggestion is, hence, that while our reasons of love are grounded in the value of individuals, the fact that we stand in a valuable loving relationship with some individuals makes a difference to how strong those reasons are. To return to the Drowning Mother case: the fact that two persons are drowning gives me a reason to save them both, but the fact that I share a loving relationship with one of them – my mother – gives added weight to my reason to save her but does not give any added weight to my reason to save the stranger. The difference between the reason I have towards my mother and the reason I have to the stranger is not one of kind but rather of degree.

It could be objected that relationships do not play the role of intensifiers at all; relationships are, the criticism may go, additional reasons. After all, it is sensible that two reasons in favor of an action also provide us with a stronger case in favor of that action than either of the two reasons taken in isolation. Moreover, we commonly talk about relationships as reasons: this is a point the advocate of the relationship view makes. While it is most likely the case that in ordinary conversations we do not distinguish between reasons, enablers, and intensifiers, this does not mean that there is no distinction to make. A reason could be an additional one if it favors doing something independently of any other reason to do that something. However, standing in a relationship with someone, regardless of how loving it is, is not an independent reason as I argued in the section on the relationship view. Relationships seem to do their normative work only when there are other reasons around. It is, therefore, more plausible to think of relationships as intensifiers rather than as additional reasons.

(Does this mean that I am committed to claiming that someone who cites the black color of shirt as a reason for buying it is wrong? Or that a person is wrong to say that the reason why they are going to an art show is because it is the last day of the exhibition? Not necessarily. After all, when we engage in ordinary conversations with others, we do so with a very different project in mind than when we engage in philosophical analyses. My friends, in all likelihood, do not particularly care, for example, whether I bought my new

shirt simply because it was black and I happen to like that color or whether the fact that it was black only gave more weight to my reasons for buying a new shirt, which were that I needed one and that this one fit me nicely. Even if I say to my friends that I bought the shirt because it was black, it is implied that I needed one and that this one looked good on me. Saying "It's black." is a shorthand because my friends, like most people, are not really interested in listening to me giving them a precise elaboration: they just want to hear something that is explanatorily relevant.)

This justification is a version of the individuals view as it begins from the value of individuals. It thus has a proper focus. However, it lacks the mysteriousness problem of the individuals view as formulated by Keller inasmuch as it does not claim that the reason I have to save my mother is different in kind from the reason I have to save a stranger. Both these reasons are of the same kind: they stem from the value of individuals. That the only reasons we have to act are grounded in the value of individuals also makes the justification an impartial one. Nevertheless, the fact that we stand in valuable loving relationships with some people but not with others, modifies our reasons by making our reason to attend to a particular someone stronger than our reason to attend to anyone. Given that this version of the individuals view makes use of valuable loving relationships too, it also circumvents the extension problem of the projects view. Incidentally, I also believe that this understanding of the role loving relationships play in making our moral decisions is "truer to the phenomenology of partiality" (Keller 2013: 80).

However, the story remains incomplete; for, why do relationships play this part? Absent an explanation, it remains unclear why reasons of love are not morally arbitrary. This brings us to the second step of a more satisfactory account of reasons of love.

The Second Step Towards a Solution: Relationships and a Meaningful Life

The second step towards a full story about reasons of love is to account for the relevance of valuable loving relationships. Why might my relationship with my mother give more weight to my reason to save her? This is the point at which, I think, the projects view (or at least what I take the lesson behind it to be) can come to the rescue, though what follows is certainly not the mainstream interpretation of the projects view.

Williams's writing is a delightful fusion of thoughtful and obscure. His idea of ground projects, examined in the previous section, is no exception. There are two ways in which one can interpret the thought that our ground project justify reasons of love. One strand is anchored in a particular view about personal identity: ground projects have their characteristic normative power because of the key role they play in constituting who we are as persons.

I examined this idea in section 2; I thus leave it aside and focus on the second strand of thought. The second element of Williams's view maintains that ground projects give meaning to our lives. Susan Wolf maintains that the idea of a meaningful life is crucial to Williams's thought. As I understand her, Wolf (2010) holds that a meaningful life, or a life worth living, is a life which consists in the pursuit of objectively valuable goods. The suggestion, then, is that loving relationships are objectively valuable goods; as such, they are one important ingredient of a life worth living. Why do valuable loving relationships figure in living a meaningful life? To answer to this question we need to see what the value of loving relationship consists in.

Loving relationships contribute to a life worth living in various ways.[11] First, being the kind of creatures that we are, it is plausible to think that we need personalized relationships in which we are valued for who we are and in which we value others for who they are in order to have a sense of belonging. This need is not only deep but it is all-encompassing too: we typically want to have many of our needs met within the context of loving relationships. We commonly prefer to eat, live, travel, learn, and play with people with whom we have, or with whom we would welcome, a loving relationship.

Second, relying on the empirical studies conducted by John T. Cacioppo and his colleagues, Kimberley Brownlee argues that "when we are deprived of adequate social connections ... we tend to break down mentally, emotionally, and physically" (2016: 55). Indeed, as the research Brownlee cities indicates, valuable loving relationships contribute to health and longevity. An actual or perceived lack of loving relationships has been linked to numerous detrimental health outcomes, such as greater likelihood of increased systolic blood pressure and cardiovascular diseases, depression and anxiety, personality disorders, impaired cognitive performance and decline of cognitive abilities over time, and increased risk of Alzheimer's disease and dementia (Cacioppo and Patrick 2008. Cited in Brownlee 2016).

Third, relationships arguably play a crucial role in the development of autonomy; alternatively, they might constitute autonomy itself. These are two claims of those who propose a relational approach to understanding individual autonomy. In any case, relational conceptions of autonomy, which stem mostly from feminist insights, stress the ubiquitous role that relationships play in a person's self-conception and which must be taken into account when outlining the conditions for individual autonomy (Mackenzie and Stoljar 2000).

Finally, it is plausible to hold that loving relationships are a necessary for achieving some other valuable goods, such as self-confidence and trust in others. Indeed, our friends, family members, and lovers provide us with necessary encouragement and advice about our life plans and about our

11 What follows is not intended to be a comprehensive explanation.

abilities to carry those plans out. (Think of the encouragement and advice you received from your parents, for example, when choosing what to study. Or think of the support you received from your partner or friends when you ventured into a risky business.) It is also within loving relationships that we develop a sense of trust in others. Having a sense of trust is not only crucial for survival (especially for those who depend on others for care: children, the elderly, and the physically and mentally impaired) (Kittay 2011) but it is also necessary in order to cooperate with others (Friedrich and Southwood 2011). The sense of trust we develop through our interpersonal relationships is germane, furthermore, for living in a political community: we need to trust fellow citizens, institutions, and politicians to uphold the social contract in order for society to function and to function well (Govier 1997, O'Neill 2002).

The argument, then, is as follows. Loving relationships contribute to a life worth living. If loving relationships are a vital ingredient for a life worth living, then they affect our reasons for action. The best explanation of how exactly loving relationships affect our reasons for action is that they act as intensifiers. That loving relationships play this role does not make reasons of love morally arbitrary because living a worthy life is morally relevant. Indeed, I can hardly think of anything that matters more to us than having a life worth living. And this matters equally to each and every one of us.

Conclusion

The main goal of this paper was to provide a compelling justification of reasons of love. To that end, I examined three prominent accounts in the literature – the projects view, the relationship view, and one version of the individuals view. These three ways of justifying reasons of love are ultimately unsuccessful, or so I argued. The projects view, first, lacks comprehensiveness; second, the relationship view does not have the proper focus; finally, the individuals view, at least in Keller's articulation, makes reasons of love not only vague but also a matter of moral chance. Nevertheless, since there is much support for the individuals view, I then presented a version of the individuals view that circumvents the problems which cause trouble for Keller's account. I suggested that while the justification of reasons of love is firmly grounded in the equal value of individual persons, the reason I have towards my loved ones has more weight because loving relationships play the role of intensifiers of reasons. Such a version of the individuals view avoids making reasons of love mysterious – for, there is only a difference in degree but not in kind – and morally arbitrary – for, loving relationships are an important part of a life worth living. If you are convinced by such a multi-layered story, you can then proceed with a clear moral conscience to throw the life jacker to your drowning mother should you ever find yourself in such a situation. Hopefully, you won't.

I promised to deliver one other thing at the beginning of the paper. Due to brevity of space and the intricacy of the issue, I can only canvass it briefly. Namely, let us accept that we are justified in giving preference to our loved ones over strangers. It surely cannot be the case that we are always morally permitted to give preference to our loved ones: benefiting our loved ones has its limits. What is that limit? This is a difficult question, for there are plenty of situations in which we might find ourselves that would fall under a gray area. Is there a principled way to distinguish between cases in which it is morally permissible to be partial to our loved ones and cases in which it is morally impermissible to be partial to our loved ones?

I think that a sensible answer to this question lies in whether the benefits we give to our loved ones are ours to give or not. If I am a public servant (national or international), for example, I am charged with considering the benefit of everyone equally (baring deontological considerations). Therefore, it would be morally impermissible to be partial to my loved ones simply because they are my loved ones: the goods I command are not mine to give away. It is the violation of this requirement, I think, which fuels the idea that practices such as nepotism and cronyism (to take one example) are morally wrong. In juxtaposition, in cases in which the goods I am allocating are mine (be they material or non-material), reasons of love are permitted, perhaps even required (Hooker 2010). To be sure, this is merely a tentative and a highly unsophisticated response. Luckily, my aim in this paper was not to settle this issue but merely to provide an account that gives a plausible story behind reasons of love.

References

Anderson, Elizabeth (1999). "What Is the Point of Equality?" *Ethics* 109 (2): 287–337.

Brownlee, Kimberley (2016). "Ethical Dilemmas of Sociability." *Utilitas* 28 (1): 54–72.

Cacioppo, John T. and Patrick, William (2008). *Loneliness: Human Nature and the Need for Social Connection*. New York: W. W. Norton & Company.

Christiano, Thomas (2006). "A Foundation for Egalitarianism." In: Nils Holtug and Kasper Lippert-Rasmussen (eds.). *Egalitarianism: New Essays on the Nature and Value of Equality*. Oxford: Clarendon Press.

Dancy, Jonathan (2004). *Ethics Without Principles*. Oxford: Oxford University Press.

Dworkin, Ronald (1977). *Taking Rights Seriously*. Cambridge, MA: Harvard University Press.

Dworkin, Ronald (2000). *Sovereign Virtue: The Theory and Practice of Equality*. Cambridge, MA: Harvard University Press.

Dworkin, Ronald (2011). *Justice for Hedgehogs*. Cambridge, MA: The Belknap Press of Harvard University Press.

Friedrich, Daniel, and Southwood, Nicholas (2011). "Promises and Trust." In: Hanoch Sheinman (ed.). *Promises and Agreements: Philosophical Essays*. Oxford: Oxford University Press.

Govier, Trudy (1997). *Social Trust and Human Communities*. Montreal: McGill-Queen's University Press.

Held, Virginia (2006). *The Ethics of Care: Personal, Political, and Global*. New York: Oxford University Press.

Hooker, Brad (2010). "When Is Impartiality Morally Appropriate?" In: Brian Feltham and John Cottingham (eds.). *Partiality and Impartiality: Morality, Special Relationships, and the Wider World.* Oxford: Oxford University Press.

Jeske, Diane (2008). *Rationality and Moral Theory: How Intimacy Generates Reasons*. New York: Routledge.

Keller, Simon (2013). *Partiality*. Princeton: Princeton University Press.

Kittay, Eva F. (2011). "The Ethics of Care, Dependence, and Disability." *Ratio Juris* 24 (1): 49–58.

Kolodny, Niko (2003). "Love as Valuing a Relationship." *The Philosophical Review* 112 (2): 135–189.

Kymlicka, Will (2002). *Contemporary Political Philosophy: An Introduction, 2nd edition*. Oxford: Oxford University Press.

Mackenzie, Catriona and Stoljar, Natalie (2000). "Introduction: Autonomy Refigured." In: Catriona Mackenzie and Natalie Stoljar. *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self*. New York: Oxford University Press.

Murdoch, Iris (1970). *The Sovereignty of Good*. London: Routledge.

O'Neill, Onora (2002). *A Question of Trust: The BBC Reith Lectures*. Cambridge: Cambridge University Press.

Scheffler, Samuel (2006). "Projects, Relationships, and Reasons." In: Michael Smith et al. (eds.). *Reason And Value: Themes from the Moral Philosophy of Joseph Raz*. Oxford: Oxford University Press: 247–269.

Scheffler, Samuel (2010). "Morality and Reasonable Partiality." In: Brian Feltham and John Cottingham (eds.). *Partiality and Impartiality: Morality, Special Relationships, and the Wider World*. Oxford: Oxford University Press: 98–130.

Williams, Bernard (1981). *Moral Luck: Philosophical Papers 1973–1980*. Cambridge: Cambridge University Press.

Wolf, Susan (1992). "Morality and Partiality." *Philosophical Perspectives* 6: 243–259.

Wolf, Susan (2010). *Meaning in Life and Why It Matters*. (With Commentary by John Koethe, Robert M. Adams, Nomy Arpaly, and Jonathan Haidt.) Princeton: Princeton University Press.

REVIEWERS FOR BELGRADE PHILOSOPHICAL ANNUAL (YEARS 2017, 2018 ND 2019):

James Connelly (Trent)
Kaja Damnjanović (Belgrade)
John Erikssson (Gothenburg)
Carrie Figdor (Iowa)
Russell Goodman (New Mexico)
Bryce Huebner (Georgetown)
Brian Huss (York)
Luka Malatesti (Rijeka)
Alex Miller (Otago)
Charles Miller (Wake Forrest)
Andrej Jandrić (Belgrade)
Leon Kojen (Belgrade)
Maja Kutlača (Osnabrük)
Miljana Milojević (Begrade)
Voin Milevski (Belgrade)
William Ramsey (Las Vegas)
Severin Schroeder (Reading)
David Stern (Iowa)
Caj Strandberg (Oslo)

CIP – Каталогизација у публикацији
Народна библиотека Србије, Београд

1

FILOZOFSKI godišnjak = Belgrade philosophical annual / editor Slobodan Perović. – God. 1, br. 1 (1988)– . – Belgrade : Institute of Philosophy, Faculty of Philosophy, 1988- (Belgrade : Službeni glasnik). – 24 cm

Godišnje. – Glavni stvarni naslov od br. 28 (2015) Belgrade philosophical annual. – Tekst na engl. jeziku.

ISSN 0353-3891 = Filozofski godišnjak

COBISS.SR-ID 15073792