Judgment and Decision Making (2024), **Vol. 19**:e17 1–29
doi:10.1017/jdm.2024.16

THEORY ARTICLE

What's moral wiggle room? A theory specification

Alina Fahrenwaldt[1,2], Fiona tho Pesch[2,3,4], Susann Fiedler[2,5] and Anna Baumert[2,4]

[1]University of Cologne, Cologne, Germany; [2]Max Planck Institute for Research on Collective Goods, Bonn, Germany; [3]École Normale Supérieure, Université PSL, Paris, France; [4]University of Wuppertal, Wuppertal, Germany and [5]Vienna University of Economics and Business, Vienna, Austria

Corresponding authors: Alina Fahrenwaldt and Fiona tho Pesch; Emails: alinafahrenwaldt@web.de; fiona@thopesch.de

The first authorship is shared by A.F. and F.t.P., as they contributed equally to the paper.

Received: 22 March 2024; **Accepted:** 2 May 2024

Keywords: moral wiggle room; theory specification; ignorance; prosocial behavior; social image; self-image; social preferences; norms; guilt

Abstract

The term 'moral wiggle room' (MWR) is often used to describe features of social situations that reduce the transparency between behaviors and their consequences. Previous research found that MWR decreases the likelihood of prosocial behavior and inferred that prosocial behavior is driven not only by genuine prosocial preferences but also by the desire to appear prosocially. Unfortunately, this postulation has never been specified as a theory. Consequently, studies testing the MWR effect reveal substantial heterogeneity in the understanding of core concepts, their operationalizations, and boundary conditions. To advance the field of MWR research, we remove these ambiguities by providing a verbal proposition-based theory specification. We first outline the original formulation of the MWR effect and its mediating mechanism, and we identify its loopholes. On this basis, we propose, refine, and distinguish between core propositions and auxiliary assumptions as well as relevant concepts and their operationalizations. The result is a fully testable theory of MWR (MWR-T) that includes a sharpened concept of MWR, distinguishes between three underlying psychological mechanisms of the behavioral MWR effect (i.e., anticipated social image damage, perceived social norms, and anticipatory guilt), and takes into account the role of individual differences in susceptibility to MWR (i.e., the joint effect of dispositional other-regarding preferences and social image concerns). Lastly, we relate MWR-T to existing theories and draw a roadmap for future work. With our contribution, we hope to stimulate more rigorous research on MWR and provide an example of the utility of verbal proposition-based theory specification.

1. Introduction

The nature of human prosocial behavior has fascinated philosophers and scientists for centuries. The most contested claim in this debate is probably the assumption that prosocial behavior (e.g., sharing, donating, and helping) is motivated by genuine prosocial preferences (i.e., people receiving a utility from behaving prosocially; Andreoni & Miller, 2002; Charness & Rabin, 2002; Fehr & Schmidt, 1999). This assumption has been challenged by experimental studies demonstrating the situational conditionality of prosocial behavior, suggesting that this behavior may depend on additional motives other than genuine prosocial preferences. Specifically, a seminal paper by Dana et al. (2007; DWK hereafter) experimented with features of a social decision situation (the *Dictator Game*; Forsythe et al.,

1994) by reducing the transparency between an agent’s behavior and the outcome for themselves and a passive recipient. They argued that this reduced transparency created a *moral wiggle room* (MWR), allowing people to behave selfishly without appearing selfish. On this basis, DWK predicted that, compared to a transparent social decision situation, the presence of MWR would reduce prosociality. Their empirical findings supported this prediction. They interpreted their findings by showing that prosociality is partly driven by a desire to *appear* prosocial rather than by genuine prosocial preferences over outcomes alone.

Although these findings and the proposed theoretical explanation are intriguing, they have never been formally specified as a theory. Accordingly, large parts of the original hypotheses and interpretations are underspecified and contain several loopholes and inconsistencies. As we demonstrate in the present paper, these problems of the original publication can be remedied. That is, by employing a verbal proposition-based theory specification, we were able to formulate an advanced, unambiguous, and testable theory of MWR (MWR-T).

In the remainder of this article, we first provide a brief outline of the existing research on MWR and explain why we need a verbal proposition-based specification. Subsequently, we utilize this method to specify DWK’s original claims about MWR and its underlying mechanisms and we point out several loopholes in their formulation. We then provide solutions to these loopholes, distinguish between core propositions and auxiliary assumptions (most of which were not made explicit in the original formulation), and clarify the core concepts and their operationalizations. The result is a fully testable MWR-T that includes a sharpened definition of MWR, distinguishes between three underlying psychological mechanisms (anticipated social image damage, perceived social norms, and anticipatory guilt), and takes into account the role of individual differences in susceptibility to MWR. We conclude with an integration of MWR theory into the broader landscape of related social science theories, summarize our key insights gained in the theory specification process, and highlight questions that should be addressed by future theoretical and empirical research.

1.1. Brief outline of research on MWR

Since its inception in 2007 (DWK), the concept of MWR has sparked enormous interest among social scientists. According to Google Scholar, by June 2024, more than 1800 papers have mentioned the term ‘moral wiggle room’, and the original paper by DWK has been cited over 1700 times. In this voluminous literature, one can identify different operationalizations of MWR, i.e., experimental manipulations used to induce MWR. The operationalization that has received the most attention is *strategic ignorance*. It describes experimental setups which allow people to avoid information on the outcome that a self-profiting decision has for the other person, liberating people to act selfishly without appearing to be selfish to others and possibly themselves (DWK; Bartling et al., 2014; Bell et al., 2017; D’Adda et al., 2018; Ehrich & Irwin, 2005; Grossman, 2014; Matthey & Regner, 2011; Momsen & Ohndorf, 2020). Another operationalization of MWR is the introduction of *uncertainty between behavior and outcome*, meaning experimental manipulations preventing the recipient from knowing whether a decision has been made by the agent or by another entity (DWK).

Furthermore, it has been proposed that the presence of *outcome risk* and *ambiguity* in a social decision situation can be exploited to behave selfishly while appearing to be risk-averse or risk-seeking (Exley, 2016). In addition, some other experimental paradigms, which have not yet explicitly been linked to the MWR concept, seem to be closely related to it. One example would be study designs introducing *information asymmetry* where the recipient of a benefit does not know the initial endowment of the benefactor and thus cannot infer with certainty whether the received outcome was fair or not (Ockenfels & Werner, 2012). Another example would be studies testing so-called *default and omission effects*, in which agents can plausibly claim to have missed the chance to choose prosocially (Gärtner & Sandberg, 2017). All of these operationalizations of MWR have in common that they reduce prosocial behavior compared to a more transparent baseline condition. For a list of the different

https://doi.org/10.1017/jdm.2024.16 Published online by Cambridge University Press

operationalizations that have been used in research on MWR, see Appendix, Table A2 (for more information, see Supplementary Material A, Table SA1, available online).

1.2. Why and how to achieve a verbal proposition-based specification of MWR theory

It could be argued that the reported diversity in operationalizations of MWR aids comprehensive theory testing by allowing MWR effects to be tested from different conceptual angles. However, a review of the literature indicates that researchers diverge not only in their operationalization of the MWR concept but also in their understanding of the concept itself. For example, while DWK stated that the label MWR describes certain situational characteristics, some researchers now employ it for any kind of justificatory cognition for immoral behavior (e.g., D'Adda et al., 2018). Heterogeneity is also present in researchers' understanding of the generality versus specificity of the MWR effect (i.e., the boundary conditions under which it is supposed to be relevant and testable). For example, some researchers generalize the MWR effect to reciprocal (e.g., van der Weele et al., 2014), to strategic (e.g., Bolton et al., 2019), or even to purely vicarious decisions (Cerrone & Engel, 2019; for a list of decision settings in which MWR was tested, see Supplementary Material A, Table SA2, available online). This heterogeneity in the understanding of a theory's central concepts, their operationalizations, and the boundary conditions of central claims is problematic because it hampers the comparability, interpretability, and replicability of research findings (Camerer et al., 2018; Smaldino, 2019) and thus scientific efficiency and progress.

As noted before, the claims of the MWR effect and its underlying mechanisms have never been formally specified as a theory and we believe that this may be the primary reason for the described heterogeneity. To achieve a consensual understanding of a scientific theory and enable stringent empirical tests, it is necessary to have available a precise formulation of all propositions and their auxiliary assumptions as well as precise concept definitions and operationalizations (e.g., Asendorpf et al., 2016), thus offering a roadmap for efficient testing (e.g., Glöckner & Betsch, 2011; Gollwitzer & Schwabe, 2020). Such a theory specification can range from a nonformal, verbal clarification of the theory to its reconstruction in a formal language (e.g., predicate logic or set theory; see, e.g., Balzer & Moulines, 1996). Although some theorists view formalization as the ultimate goal of theory specification, we believe that an intermediate solution based on a systematic procedure and clear rules while retaining real language (instead of converting propositions into symbols) is a more accessible form of theory specification. Consequently, it may be more likely to be read and understood by many researchers in the relevant research field.

Glöckner and Betsch (2011) introduced standards for such a verbal proposition-based theory specification. Their approach aims to increase the *empirical content* of theories (i.e., increasing both a theory's universality and precision in terms of the clarity of predictions and avoidance of contradictions and tautologies; Popper, 1934). Glöckner and Betsch suggest that a well-specified theory should consist of a finite set of clear-cut *propositions* expressing relationships or causal effects among concepts of the theory, which together fully describe the theory. When specifying unidirectional causal effects, a *proposition* consists of two elements: *antecedent* and *consequence* (written as unambiguous if-then statements). If a theory includes mediating mechanisms, they should be presented as multiple (interconnected) propositions. If the antecedent or consequence of a proposition includes multiple *concepts*, they are linked through logical (AND, OR, etc.) operations. Crucially, all concepts appearing in the set of a theory's propositions have to be defined in an unambiguous and testable manner. A good theory specification should also describe how to measure and manipulate these concepts; i.e., it should outline their *operationalization*. Lastly, the boundary conditions for the theory should be explicitly stated as *auxiliary assumptions*. These auxiliary assumptions hold information regarding the subgroup of people or situations the theory applies to, thus concerning the trade-off between a theory's *specificity and generalizability.* They are also necessary to isolate the effects of interest and rule out potentially confounding factors. Lastly, a good theory specification should identify critical properties allowing for theory falsification.

https://doi.org/10.1017/jdm.2024.16 Published online by Cambridge University Press

Table 1. *Propositions derived from the original paper on MWR (DWK).*

Proposition	Antecedent	Consequence
For all agents and situations specified in the respective auxiliary assumptions:		
1	IF MWR (instead of no MWR)	THEN higher likelihood of selfish behavior
2a	IF MWR (instead of no MWR)	THEN reduced relevance of fairness norms and constraints (i.e., feeling less compelled to give or having an excuse or justification not to give)
2b	IF reduced relevance of fairness norms and constraints (i.e., feeling less compelled to give or having an excuse or justification not to give)	THEN higher likelihood of selfish behavior

Note: The formulations in this table are abbreviations of full sentences. For example, Proposition 1 reads as '*If a person is in a situation containing moral wiggle room, then this person will be more likely to show selfish behavior relative to a situation not containing MWR'.* Furthermore, in DWK's study designs, all proposed links between concepts appear to be viewed as probabilistic rather than deterministic and as linear relationships.

We believe that a verbal proposition-based specification of a MWR-T, following the standards proposed by Glöckner and Betsch (2011), will reduce existing ambiguities and prevent future misunderstandings of the theory. Moreover, it will serve to identify the most relevant open questions that should be addressed by future research.

2. Specification of the original formulation of MWR effects

The original paper on MWR (DWK) contained only a rudimental formulation of DWK's theoretical assumptions. Although a partial specification of propositions and core concepts can be deduced if one also considers the study designs employed by DWK (for an overview, see Supplementary Material B, available online, the original paper left critical aspects underspecified. Therefore, we present the specification of the original formulation directly together with important loopholes. How these problems can be amended is described in the subsequent section of this article, where we will present a fully specified MWR-T.

2.1. Propositions

At the heart of the original formulation (DWK) lies the proposition that in situations containing MWR (compared to situations not containing MWR) people are more likely to show selfish behavior (see Table 1, Proposition 1).[1]

DWK explain this behavioral effect of MWR (Propositions 2a and 2b) by MWR offering agents an 'excuse' or 'justification' not to give (DWK, p. 69) or the opportunity of '*feeling [less] compelled to give'* (DWK, p. 77–78). The authors use these formulations interchangeably with a third claim that agents perceive a change in '*norms and constraints'* (DWK, p. 78) in MWR situations. Specifically, DWK propose that the norm of fairness is perceived as less relevant (i.e., as 'less binding' or as 'compet[ing] with other norms', DWK, p. 78) in situations with MWR. The authors additionally state that other mechanisms could underlie the MWR effect, but they do not specify them.

[1] Note that this proposition *could* be called the central empirical claim of the specification but is included as a proposition since most research on MWR is driven not only by explaining the effect of MWR on behavior but also by testing this behavioral effect with different operationalizations and in different decision situations.

https://doi.org/10.1017/jdm.2024.16 Published online by Cambridge University Press

2.2. Concept definitions and operationalizations

2.2.1. MWR

According to DWK, MWR is defined as situational characteristics that remove the transparency between (selfish) behavior and outcomes. 'Transparency' here refers to the 'commonly known one-to-one mapping between the [agent's] actions and the outcomes to both parties' (DWK, p.69). DWK operationalized (induced) MWR in three different ways: by implementing experimental treatments termed (i) 'hidden information', (ii) 'plausible deniability', and (iii) 'multiple dictator' (see Supplementary Material B available online for a detailed description of the three original treatments). No MWR describes situations with full transparency between behavior and outcomes (i.e., the baseline setting).

➔ **Loophole 1: Inconsistency between the definition of MWR and the proposed explanatory mechanism**
According to the original definition of MWR, the presence of MWR makes it more difficult to infer whether an observed *outcome* resulted from the agent's *behavior*, i.e., whether the agent caused, and thus can be held accountable for, the outcome. At the same time, DWK propose that a change in norms and constraints drives the effect of MWR on an agent's choice (Propositions 2a and 2b). However, reduced accountability does not imply changed norms and constraints. Moreover, a lack of accountability alone can liberate the agent to behave selfishly, even if the situational norms and constraints remain unchanged (Krysowski & Tremewan, 2021). Therefore, the proposed explanation of the MWR effect in terms of changed norms makes more sense if the concept of MWR is redefined as situational characteristics which make it difficult to infer an agent's *intentions* behind a behavior. Specifically, such situational characteristics could allow for attributing the behavior to other plausible reasons (e.g., being short on time, not wanting to be nosy, being overwhelmed by the decision). Consequently, in these situations, observers of the agent's behavior, as well as the agent themselves, may perceive a change in the relevant social norms such that selfish behavior is seen as less socially inappropriate.

➔ **Loophole 2: Unsuitable operationalizations of the concept of MWR**
Of the three experimental treatments used by DWK to induce MWR, the '*multiple dictators'* treatment does not fit well with DWK's definition of MWR. In this treatment, the prosocial option for a passive recipient is implemented if at least one member of a group of agents chooses this option (over a selfish option, i.e., an agent-profiting but recipient-disadvantaging outcome). In this situation, transparency exists only in the case of a prosocial outcome – if this outcome occurs, the recipient can infer only that at least one agent acted prosocially, but not who they were. In contrast, if the selfish outcome occurs, the recipient can infer that all agents chose the selfish option. Moreover, in this case, the agent's intentions are also directly inferable. Because of this, increases in selfish behavior observed in this treatment cannot easily be explained by MWR effects; they are more likely due to a diffusion of responsibility (Darley & Latané, 1968).

2.2.2. Selfish behavior

Following DWK (and a general understanding of economics), selfish behavior is defined as behavior maximizing one's own profit while disregarding other people's payoff. DWK operationalize this concept as a binary choice between a selfish option profiting the agent more and the recipient less and a second option which is more egalitarian. Selfish behavior is then operationally defined as choosing the selfish option. DWK estimate the effect of MWR on selfish behavior at the population level.

2.2.3. Relevant mechanism

The original publication on MWR did not contain a clear description of the psychological mechanisms proposed to mediate the MWR effect nor did it contain precise definitions and operationalizations of the concepts 'norms and constraints', 'excuse', 'justification', and 'feeling [less] compelled to give' (DWK, p. 69 & p. 78). Following the general logic of DWK's experiments, concepts of the mechanism propositions are likely estimated at the population level.

https://doi.org/10.1017/jdm.2024.16 Published online by Cambridge University Press

➔ **Loophole 3: Lack of differentiation of psychological mechanisms**
According to Popper (1934), the hallmark of an empirical theory is that it can be falsified by evidence. To allow for the falsification of proposed mechanisms mediating the behavioral MWR effect, there needs to be clarity about these underlying psychological mechanisms. As mentioned before (Propositions 2a and 2b, Table 1), DWK propose that MWR reduces prosocial behavior because it provides the agent with 'excuse[s]' or 'justifications' for not giving and that agents 'do not feel compelled to give' in situations with MWR (DWK, p. 69 & p. 78). As also noted, DWK use this proposition interchangeably with the proposition that MWR changes the (bindingness or availability of competing) norms and constraints (Propositions 2a and 2b, Table 1). However, psychologically, perceptions of norms are not the same as feelings. Moreover, DWK state that the proposed psychological mechanism may only be 'one way' to account for their results (DWK, p. 78). This statement is problematic because allowing for other, unspecified mediating mechanisms runs the risk of making a theory unfalsifiable (see Glöckner & Betsch, 2011).

➔ **Loophole 4: Lack of clear definition and operationalization of mechanism concepts**
To allow for strong tests of the mechanism propositions, one needs to clarify the definitions and valid operationalizations of the concepts appearing in these propositions. For the proposed *change in norms and constraints*, this can be accomplished by specifying whether MWR affects behavior through an objective change in the prevailing social norms (i.e., changes in what most people find appropriate) or through a change in an agent's perception of these prevailing norms. Similarly, one needs to clarify what *feeling (less) compelled to give* means and how this feeling can be measured. Lastly, one needs to specify whether MWR provides agents with *justifications* for their selfishness in front of others (social image), in front of themselves (self-image), or both.

2.3. Auxiliary assumptions

In the original paper (DWK), no auxiliary assumptions of the authors' propositions were explicitly mentioned. Nevertheless, four auxiliary assumptions regarding the social decision situations in which the proposed effects are observable can be derived from DWK's descriptions of the experimental setups designed to test their propositions and their additional explanations (for their necessity for the different propositions, see Supplementary Material B, Table SB3, available online). These auxiliary assumptions are:

1. The decision must have consequences for oneself and others, and the interests of these parties must conflict.
2. The presence of MWR in the social decision situation must not restrict the agent's choice (i.e., their ability to implement any of the outcomes available without MWR).
3. The interaction between the agent and the recipient is nonstrategic and unilateral.[2]
4. There are two independent and sometimes conflicting motives active in agents in the population: (a) an agent's preferences over payoff distributions and (b) an agent's self- or social image concerns.[3]

➔ **Loophole 5: Lack of clear definitions and operationalizations of agents' motivations**
In our view, the concept of *image concerns* (Auxiliary Assumption 4) is not sufficiently specified. That is, in their original formulation, DWK are ambiguous about the relevance of agents' concerns about their social image versus their self-image for the proposed effects. On the one hand, MWR

[2] Note that this excludes social situations where agents and recipients interact repeatedly. In these situations, agents could be influenced by additional concerns (such as the fear of retaliation by the recipient in case of an unfair outcome). Consequently, the proposed driver of the MWR effect (Propositions 2a and 2b) would not be clearly separable (in behavior) from these additional concerns. DWK emphasize that they chose to test the effect of MWR in settings which preclude such concerns.

[3] Note that DWK did not use any specific term for this concept. We believe that the term 'image concerns', which is common in the literature on moral and social behavior, captures best what DWK mean when they state that many people 'do not want to appear selfish' (p. 68). Some readers may be more familiar with the term 'signaling', which we would like to treat as a synonym for 'image concerns'.

https://doi.org/10.1017/jdm.2024.16 Published online by Cambridge University Press

is conceptualized as a reduction in the '*commonly* known one-to-one-mapping' (DWK, p. 69, emphasis added) between behavior and outcome, suggesting that agents could be concerned with their *social* image. On the other hand, DWK assume that people 'do not want to appear selfish - *either to themselves* or others' (DWK, p. 68, emphasis added), suggesting that agents want to protect their social image, their self-image, or both. To ensure a correct understanding of relevant boundary conditions, one needs to specify the exact type of image concerns necessary for the proposed effects. Additionally, one needs to specify how these image concerns should be measured. Clarifying the relevance of these different concerns also has consequences for the definition of MWR (i.e., one needs to specify whether MWR treatments create intransparency for others, for oneself, or both).

➔ **Loophole 6: No specification of individual differences**
From Auxiliary Assumption 4, which specifies two potentially conflicting motives necessary for the proposed effects, we can also derive that the effects of MWR may depend on the relative strength of these motives in a particular agent. In other words, there could be differential effects of MWR on social behavior. Supporting this idea, the original studies by DWK found indirect evidence for such individual heterogeneity in the MWR effect.[4] Ideally, these individual differences should be incorporated into a fully specified MWR-T.

➔ **Loophole 7: No explicit specification of the response format and action space available to agents**
The original formulation of the behavioral MWR effect leaves unclear whether it is expected to occur only in binary decision situations (pitting a selfish outcome against a fair outcome) or also in situations in which agents can choose between more than two actions, perhaps even including giving that exceeds fairness.

3. Theory of MWR (MWR-T)

Incorporating solutions to the identified loopholes, we propose the following fully specified Theory of MWR (MWR-T; for a visualization, see Figure 1). We provide an overview of the loopholes and their solutions in Supplementary Material B, Table SB4 (available online). Lists of all propositions and auxiliary assumptions of MWR-T as well as precise definitions and operationalizations of the theory's concepts (resolving Loopholes 2, 4, and 5) can be found in the Appendix (Tables A1, A2, and A3).

3.1. Behavioral MWR effect and underlying psychological mechanisms

3.1.1. Propositions

The behavioral proposition remains essentially the same as in the original formulation, apart from describing the change in selfish behavior as an 'increase' (Proposition 1, Table 2) instead of containing it to a 'higher likelihood' (Proposition 1, Table 1). This was done to make MWR-T applicable to a broader set of dependent variables (see also the new definition of selfish behavior below). To remedy the underspecification of the psychological mechanisms proposed to underlie the MWR effect (resolving Loophole 3), MWR-T distinguishes between three conceptually distinct, but causally related, mediating mechanisms. These are (a) an anticipated damage to one's social image, (b) a change in perceived social norms, and (c) an anticipatory emotional reaction (anticipatory guilt) to the situation (see Figure 1 and Table 2).

3.1.1.1. Image mechanism: Reduction in the anticipated damage to the agent's social image

MWR-T holds that the main psychological mechanism underlying the behavioral MWR effect is a reduction in the damage which agents anticipate for their *social* image in case of selfish behavior under MWR versus no MWR (see Table 2, Propositions 2a and 2b, resolving Loopholes 4 and 5).

[4] In DWK *even without* any MWR (i.e., in transparent situations), roughly one-quarter of participants behaved selfishly, and *even with* MWR, roughly one-third of their participants decided against the selfish option. Therefore, only a fraction of people were affected by MWR.

https://doi.org/10.1017/jdm.2024.16 Published online by Cambridge University Press

Table 2. *Main propositions of MWR-T.*

Proposition	Antecedent	Consequence
For all agents and situations specified in the respective auxiliary assumptions:		
1	IF MWR (instead of no MWR)	THEN increase in selfish behavior between these situations
2a	IF MWR (instead of no MWR)	THEN decrease in anticipated social image damage attached to selfish behavior between these situations
2b	IF decrease in anticipated social image damage attached to selfish behavior between situations	THEN increase in selfish behavior between these situations
3a	IF MWR (instead of no MWR)	THEN change in perceived social norms (decrease in the perception of selfish behavior as socially inappropriate) between these situations
3b	IF change in perceived social norms (decrease in the perception of selfish behavior as socially inappropriate) between situations	THEN increase in selfish behavior between these situations
4a	IF MWR (instead of no MWR)	THEN decrease in anticipatory guilt between these situations
4b	IF decrease in anticipatory guilt between situations	THEN increase in selfish behavior between these situations
5a	IF change in perceived social norms (decrease in the perception of selfish behavior as socially inappropriate) between situations	THEN decrease in anticipated social image damage attached to selfish behavior between these situations
5b	IF decrease in anticipated social image damage attached to selfish behavior between situations	THEN decrease in anticipatory guilt between these situations

Note: The formulations in this table are abbreviations of full sentences. For example, Proposition 1 reads as '*If a person is in a situation containing moral wiggle room, then this person will show more selfish behavior relative to a situation not containing MWR'.* All proposed links should be viewed as probabilistic rather than deterministic. Furthermore, for simplicity, all relationships listed here are conceptualized to be of linear form.

This proposition aligns with other researchers' suggestions that agents might factor in anticipated social image damage when deciding whether to behave selfishly (Exley, 2016). Moreover, agents' anticipation of reduced social image damage in case of selfish behavior under MWR appears to be quite realistic as indicated by others' rating of agents' character (e.g., Grossman & van der Weele, 2017). Consequently, individuals in MWR situations behave more in line with their actual (selfish) preferences (DWK).

3.1.1.2. Normative mechanism: Change in perceived social norms

Similarly to the original formulation by DWK, we propose in MWR-T that MWR may also affect behavior by changing perceived social norms, such that selfish behavior is perceived as less socially inappropriate (see Table 2, Propositions 3a and 3b). Social norms reflect a group's (implicit) rules and standards for appropriate behavior in different situations and are often used as heuristics for decision-making (Bicchieri, 2005). The proposition that MWR makes selfish behavior socially more acceptable is supported by previous research findings. For example, choosing the self-profiting option after deciding to ignore relevant information about others' payoffs is perceived as less socially inappropriate by third-party observers (Krupka & Weber, 2013) as well as by recipients (Bolton et al., 2019; Grossman

https://doi.org/10.1017/jdm.2024.16 Published online by Cambridge University Press

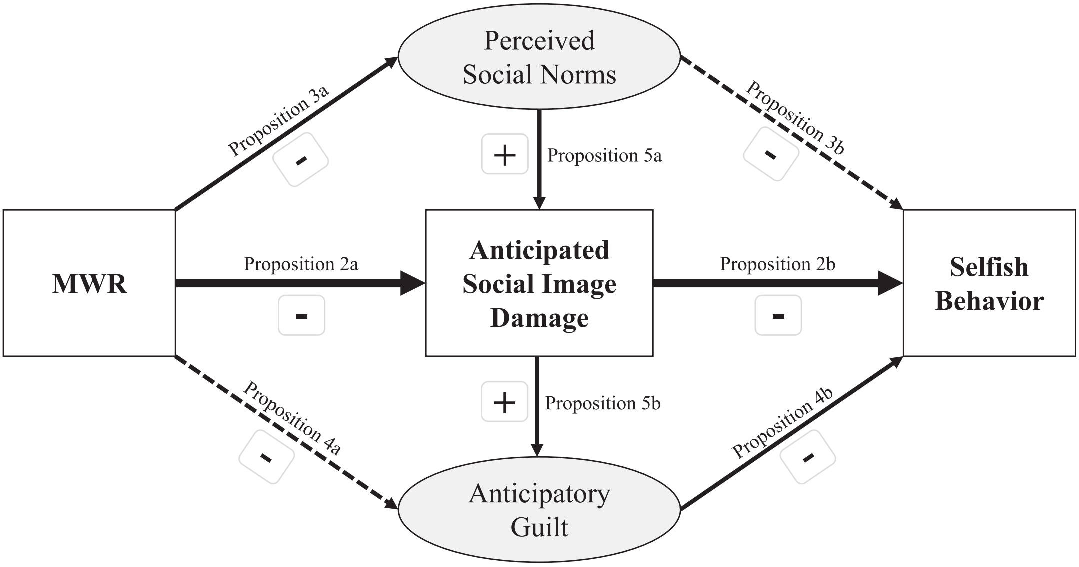

Figure 1. *Proposed psychological mechanisms underlying the effect of MWR on social behavior.*
Note: Arrows symbolize causal links. Propositions are named next to their corresponding arrows. The line style of arrows indicates the proposed importance or strength of the causal influences (bold line = most important mediating process; narrow unbroken line = causal effects of secondary importance; narrow broken line = least important causal influences when considering respective other pathways). The valence (positive versus negative) of each proposed causal relationship is indicated next to the arrows by '+' and '–', respectively. The final outcome for selfish behavior depends on the sum of valences of the respective pathways (e.g., an increase in selfish behavior resulting from the combination of Propositions 2a and 2b).

& van der Weele, 2017) compared to knowingly choosing the selfish option. Moreover, such selfishness in MWR situations results in fewer ultimatum game rejections (Conrads & Irlenbusch, 2013) and lower third-party punishment (Bartling et al., 2014).

3.1.1.3. Emotional mechanism: Reduction in anticipatory guilt

Third, MWR-T holds that MWR may also influence selfish behavior by reducing anticipatory guilt (Table 2, Propositions 4a and 4b). Anticipatory guilt is typically evoked by contemplating acts that violate sociomoral standards and has been shown to inhibit selfish and immoral behavior (Tangney et al., 2007). Indeed, some authors have suggested an anticipatory feeling of guilt for selfish behavior in transparent giving situations (i.e., situations without MWR; Feiler, 2014; Garcia et al., 2020). Vice versa, the presence of MWR may reduce selfishness-related anticipatory guilt and, consequently, increase selfish behavior (e.g., Thunström et al., 2014).

3.1.1.4. Interrelations of the three psychological mechanisms

To specify the mediating processes even further in MWR-T, we propose that the change in anticipated social image damage is the main psychological mechanism by which MWR affects behavior, whereas the other two psychological mechanisms are of secondary importance. In addition, we posit that the three proposed processes are interrelated (see Figure 1 and Table 2, Propositions 5a and 5b).

Normative mechanism and image mechanism. Adherence to social norms may be used as a signal of one's own morality, in an attempt to create a positive social image (Andreoni & Bernheim, 2009). Conversely, not adhering to social norms may result in anticipated image damage. Therefore, we propose that a MWR-induced change in perceived social norms also reduces anticipated image damage.

Emotional mechanism and image mechanism. Guilt has been associated with one's anticipated social image (Larson & Capra, 2009). For MWR-T, we propose that reductions in anticipatory guilt may result

https://doi.org/10.1017/jdm.2024.16 Published online by Cambridge University Press

from reduced anticipated social image damage. This is plausible because anticipatory guilt is thought to result from the appraisal of immoral actions attributed to oneself.[5]

3.1.1.5. Falsification of the theory

The behavioral proposition of MWR (Proposition 1 in Table 2) in MWR-T can be tested independently of the propositions about mediating mechanisms (the remaining propositions in Table 2) and does not require that the mechanism propositions be confirmed. However, if supporting evidence for the behavioral proposition (Proposition 1) is not accompanied by evidence for at least one of the three psychological mechanisms (Propositions 2a-4b), this would constitute a falsification of MWR-T. The interrelation propositions (Propositions 5a and 5b) are only possible theoretical refinements of the other propositions. A refutation of these propositions alone should therefore not be regarded as a falsification of the whole theory.

3.1.2. Concept definitions and operationalizations

For a list of possible operationalizations of all concepts used in MWR-T, see the Appendix, Table A2. Unsuitable operationalizations were excluded (e.g., the '*multiple dictators'* treatment for inducing MWR, resolving Loophole 2). In the following, we provide clear and consistent definitions for all included concepts (resolving Loopholes 1, 4, and 5).

Note that the effects of MWR are assumed to exist at the individual level, but all links between concepts are viewed as probabilistic (allowing for systematic and unsystematic variance). Therefore, all effects proposed in Table 2 are estimated at the group level.

3.1.2.1. MWR

In MWR-T, we redefine MWR as situational characteristics that obfuscate the signal which the outcome of an own-payoff-maximizing (i.e., potentially selfish) behavior sends to others about one's intention to behave selfishly. Thus, the focus lies on whether the agent's selfish *intention* can be inferred from the observable outcome and not whether the agent's *behavior* can be observed or inferred from the outcome. Consequently, no MWR describes situations with full transparency between intention and outcome (i.e., the baseline setting).[6]

Note that this redefinition of MWR includes intransparency between behavior and outcome: If an agent's behavior is unknown, the intention behind it cannot be inferred either. However, with this redefinition, MWR can exist (and take effect) even when the agent's behavior is observable. This new definition is more consistent with the proposed psychological mechanisms, such as a change in norms and constraints (resolving Loophole 1). It is also supported by empirical findings. For example, third parties punished an agent's own payoff-maximizing decisions less, even if they could observe these decisions, as long as the agent had not revealed the recipient's outcome prior to their decision (Bartling et al., 2014). In other words, even in situations in which agents could be held accountable for their *behavior*, others still punished them less if they were not sure about the *intentions* behind this behavior.

3.1.2.2. Selfish behavior

We adapt the definition from the original formulations of DWK for MWR-T by defining *selfish behavior* as choosing a (more) selfish distribution option over one or several (more) prosocial distribution option(s). This definition allows to operationalize selfish behavior as a binary, ordinal, or continuous variable at the individual level. A (more) selfish option always yields a higher payoff for the agent and a lower payoff for the recipient, whereas (more) prosocial options are less profitable for the agent but reduce the outcome inequality between the agent and recipient. Note that an increase in selfish behavior in MWR experiments is usually estimated as an increase either in the likelihood of

[5] Note that this form of guilt is more in line with the 'guilt-from-disapproval' rather than the 'guilt-from-disappointment' type that has been outlined by Hauge (2016).

[6] Note that the term *outcome* refers to the overall outcome to all affected parties (i.e., the resulting distribution of resources).

https://doi.org/10.1017/jdm.2024.16 Published online by Cambridge University Press

selfish behavior (estimated from choice frequencies) or in the average degree of selfish behavior (for continuous operationalizations).

3.1.2.3. Anticipated social image damage

This concept, which we have added to MWR-T, describes agents' expectations of how negatively others will judge them in terms of their morality due to their (selfish) decisions. We assume that this expectation is generated by a cognitive appraisal process taking place when agents consider the different distribution options before actually deciding.

3.1.2.4. Perceived social norms

Social norms describe behavioral rules based on social consensus regarding appropriate behavior in a specific situation. We assume that any objective change in social norms can only be behaviorally relevant for agents if they perceive the normative change. Therefore, we propose to measure the perceived (change in) social norms (for a discussion of the difference between objective and perceived social norms, see Tankard & Paluck, 2016).

3.1.2.5. Anticipatory guilt

Anticipatory guilt is a negative emotion experienced when contemplating decision options that violate perceived moral standards. In situations, in which the relevant decision pits own-payoff-maximization against fairness considerations, anticipatory guilt is expected to arise when the agent contemplates acting selfishly.

3.1.3. Auxiliary assumptions

Auxiliary Assumptions 1–3 from the specification of DWK's original formulations remain valid. To address Loopholes 5 and 7, we extend and revise the remaining auxiliary assumptions. Note that the different auxiliary assumptions are not all necessary for each proposition (for the complete list of auxiliary assumptions, their application, and specific reasons, see the Appendix, Table A3).

First, resolving Loophole 7, we specify adequate response options. We do not restrict the response *format* to binary decisions. However, we still limit the behavioral *space* to decisions ranging from sharing nothing to sharing 50% of the resources distributable by the agent. This decision is motivated by research which found a monotone increase in appropriateness ratings within this behavioral range, but a flattening or even reversed relationship between the amount of shared resources and its rated appropriateness for sharings exceeding 50% (Krupka & Weber, 2013). Thus, giving above the 50:50 split may be motivated by very different reasons than giving less than or exactly 50 percent, which implies possible confounds. Therefore, we refrain from making predictions for the behavioral range above the 50:50 split.

4. The behavioral space should range from maximally selfish distribution options (i.e., keeping the whole initial endowment) to egalitarian distribution options (i.e., 50:50 split).

In order to specify the role of social image concerns (resolving Loophole 5), Auxiliary Assumption 4 from the original formulation (now Auxiliary Assumption 5, see below) is revised. In MWR-T, we assume that *social* image concerns are decisive for the MWR effects. Empirical results support this assumption (Andreoni & Bernheim, 2009). In contrast, self-image concerns seem to have little or no effect on behavior in MWR settings (Grossman, 2015) and may be more relevant for positive deviations from existing prosocial behavior than for the shift from selfish to prosocial behavior (Bénabou & Tirole, 2006; Bodner & Prelec, 2003; Grossman, 2015; Lazear et al., 2012). The latter behavior is not at the focus of MWR-T and may be driven by other mechanisms than those specified in MWR-T. We believe that the behavioral effect of MWR and all three psychological mechanisms in MWR-T require agents to have (a minimum level of) social image concerns.

https://doi.org/10.1017/jdm.2024.16 Published online by Cambridge University Press

Table 3. *Individual differences propositions of MWR-T.*

Proposition	Antecedent	Consequence
For all agents and situations specified in the auxiliary assumptions:		
6	THE HIGHER the social image concerns	THE GREATER the behavioral effect of MWR
7	THE MORE extreme (i.e., selfish or prosocial) the other-regarding preferences	THE SMALLER the relevance of social image concerns for the behavioral effect of MWR AND THE SMALLER the behavioral effect of MWR

Note: The formulations in this table are abbreviations of full sentences. For example, Proposition 6 reads as '*The higher a person's (dispositional) social image concerns, the greater will be the effect of that person being in a situation containing MWR compared to a situation not containing MWR on the person's behavior*'. All proposed links should be viewed as probabilistic rather than deterministic. Furthermore, for simplicity, the relationship in Proposition 6 is conceptualized to be of linear form, while Proposition 7 explicitly specifies an inverse U-shaped relationship.

5. There are two independent and sometimes conflicting motives active in agents in the population: (a) an agent's preferences over payoff distributions (henceforth *other-regarding preferences*) and (b) an agent's *social image concerns*. Additional motives possibly influencing agents in a social decision situation must be controlled for or held constant.

3.2. Accounting for heterogeneity

3.2.1. Propositions

To account for heterogeneity in the agents' preferences or motives (resolving Loophole 6), we posit in MWR-T that the effect of MWR depends on relatively stable individual differences (see Table 3) in *social image concerns* and *other-regarding preferences*.

3.2.1.1. Social image concerns

According to Auxiliary Assumption 5, the MWR effects should only be observable in a population with social image concerns. Here, we further specify that the behavioral MWR effect increases with an agent's dispositional social image concerns (Proposition 6). This proposition is motivated by the finding that social image concerns are positively related to prosocial behavior (Gotowiec & van Mastrigt, 2019), especially if the prosocial behavior is visible or not incentivized (Müller & Moshagen, 2019; Winterich et al., 2013), and that misrepresenting one's own other-regarding preferences in transparent (but not intransparent) situations seems to be driven by image concerns (Friedrichsen & Engelmann, 2013). Most likely the importance of the three psychological mechanisms also increases with the agent's dispositional social image concerns. This assumption is supported by research associating image concerns with norm adherence (Gross & Vostroknutov, 2022), anticipated image damage (Bursztyn & Jensen, 2017), and guilt proneness (Regner, 2021). However, for MWR-T we only specify the relevance of image concerns for the behavioral effect. We propose that agents' other-regarding preferences further moderate the interactive effect of social image concerns and MWR on social behavior.

3.2.1.2. Other-regarding preferences

Agents also differ in the degree to which they generally care about fairness and prosociality (Murphy et al., 2011). We propose a nonlinear effect of dispositional other-regarding preferences on social behavior for all individuals who have social image concerns (see Proposition 7): For individuals with a strong inclination to act prosocially or selfishly, the presence of MWR (and the strength of social image concerns) should be less important for their social behavior than for those individuals with more intermediate other-regarding preferences (Grossman & van der Weele, 2017).

To summarize, MWR-T holds that the two dispositional motives, social image concerns and other-regarding preferences, interact, resulting in the most pronounced behavioral MWR effect in people with strong social image concerns and moderate other-regarding preferences.

3.2.1.3. Falsification of the theory

We do not consider Propositions 6 and 7 to be core to MWR-T, but rather as an extension of the theory. Therefore, failure to support these propositions does not falsify the theory.

3.2.2. Definitions and operationalizations of the concepts appearing in Propositions 6 and 7

For a list of possible operationalizations of the two concepts, see the Appendix, Table A2. Note that both concepts are continuous rather than binary constructs. Note that all links between concepts in Propositions 6 and 7 are viewed as probabilistic (allowing for systematic and unsystematic variance). Therefore, all effects proposed in Table 3 are estimated at the group level.

3.2.2.1. Social image concerns

Social image concerns describe how much an agent cares about being evaluated positively by others. A person's social image concerns in a specific situation are assumed to be determined by the person's dispositional social image concerns (i.e., how much the person generally cares about being positively evaluated) and by domain-specific factors (e.g., whether the relevant others are their own family versus anonymous others).

3.2.2.2. Other-regarding preferences

Other-regarding preferences refer to the true preferences of an agent over the distribution of resources between themselves and a recipient. In MWR-T, we assume that these preferences are independent of social image concerns and not influenced by MWR. An agent's other-regarding preferences in a specific situation are assumed to be determined by their dispositional other-regarding preferences (i.e., how much they generally care about themselves versus others) as well as by domain-specific factors (e.g., whether the recipients are children, animals, or the environment; or whether the shared good is money, time, etc.).

3.2.3. Auxiliary assumptions

Auxiliary assumptions 1-5 specified above for MWR-T remain valid for Propositions 6 and 7 (see the Appendix, Table A3).

4. Discussion

Since its introduction into the literature by Dana, Weber, and Kuang in 2007 (DWK), researchers from all over the world have documented the effect of MWR on prosocial behavior in a multitude of settings. Closer examination reveals that these researchers differ in their understanding of the core propositions and auxiliary assumptions of the original formulation of the behavioral effect and its psychological mechanisms as well as the relevant concepts and their operationalizations. The reason for this diversity could be that DWK did not sufficiently specify their claims and explanations. This is problematic for the interpretability and comparability of empirical findings and hinders scientific progress. In the present article, we set out to remedy this shortcoming in the literature by providing a verbal proposition-based theory specification of DWK's formulation, following standards developed by Glöckner and Betsch (2011). This approach revealed several issues and loopholes in the original formulation of the effect of MWR on behavior and its mediating psychological mechanisms. By incorporating plausible solutions to the identified problems, we then proposed a fully specified MWR-T that has higher empirical content (Popper, 1934) than the original formulations by DWK and, therefore, enables strict empirical tests.

https://doi.org/10.1017/jdm.2024.16 Published online by Cambridge University Press

4.1. Key aspects of MWR-T

The most important loopholes of the original formulation by DWK concern (1) the definition of MWR, (2) the specification of underlying psychological mechanisms and their interrelation, and (3) the role of individual differences in the MWR effect. Filling these gaps through specification, we provide:

(1) A redefinition of MWR as situational characteristics that obfuscate the signal which the outcome of an own-payoff-maximizing behavior sends to others about one's intention to behave selfishly. This redefinition of MWR (a) accounts for the finding that MWR effects can be shown even in settings where an agent's behavior is observable and (b) makes the propositions of psychological mechanisms underlying the MWR effect (Propositions 2a–4b) more plausible.

(2) A disentanglement of three potential psychological mechanisms underlying the behavioral MWR effect as conceptually different, yet interrelated processes: (a) the anticipation of less damage to the agent's social image in case of acting selfishly; (b) a change in perceived social norms regarding selfishness; and (c) reduced anticipatory guilt when contemplating selfish behavior. In addition, we propose that anticipated social image damage is the most important mediating process, receiving input from social norm perceptions and providing input to anticipatory guilt. The distinction between the three psychological mechanisms that underlie the MWR effect and the specification of their possible interrelations in MWR-T provides the basis for more rigorous tests of the theory in future work.

(3) Additional propositions specifying the role and interaction of individual differences in other-regarding preferences and social image concerns. In MWR-T, we propose that the effect of MWR on social behavior generally increases with the degree of dispositional social image concerns, but that this interactive effect is weaker for agents with more extreme other-regarding preferences (i.e., agents who have very prosocial or very selfish preferences). Notably, this three-way interaction sets MWR-T apart from earlier work that considered only additive effects of other-regarding preferences and image concerns (e.g., Andreoni & Bernheim, 2009). Note that our more complex interactive proposal, if supported by empirical studies, has significant implications for research and applications (e.g., intervention designs tailored to the multi-trait personality of social agents).

Based on the new specifications of MWR-T, we provide a list of suitable manipulations of MWR to be used in future studies (Appendix, Table A2). We invite other researchers to test and potentially falsify the different elements of MWR-T (i.e., propositions, concept definitions, and operationalizations as well as auxiliary assumptions) and thereby contribute to the further development (modification, revision, extension) of the theory.

4.2. Empirical roadmap

MWR-T as specified in this article needs to be rigorously tested. This includes isolating the psychological mechanisms underlying MWR, quantifying their unique contributions and interdependencies, and testing the proposed moderating role of individual differences in other-regarding preferences and social image concerns on the MWR effect.

4.2.1. Open questions possibly requiring theory revision

There are some persistent open questions, which, once answered, may demand theory revision:

(1) Is self- or social image damage more decisive for the effect of MWR? This debate becomes more understandable when taking a closer look at the employed MWR operationalizations. For instance, in the 'willful ignorance' treatment, recipients are denied any information about whether or not the agent revealed the full outcome matrix. This protects agents' social image, but it does not explain why a substantial fraction of agents still avoid revealing the outcome information (DWK; Grossman & van der Weele, 2017). The latter can be explained by self-image protection, which is why some researchers highlight its importance for the MWR effect (Grossman & van der Weele, 2017;

https://doi.org/10.1017/jdm.2024.16 Published online by Cambridge University Press

Matthey & Regner, 2014). It is, however, not clear yet how exactly people can fool themselves into thinking that they would be less blameworthy (in front of themselves) for selfish behavior under MWR (see Grossman, 2010). Other research points toward an alternative possibility, namely that agents feel observed and apply perceived social norms even in the absence of actual observation (i.e., 'internalized social image'; tho Pesch & Dana, 2024). This may depend on other cues of lacking anonymity (Haley & Fessler, 2005) and individuals' propensity to use heuristics (Jordan & Rand, 2020). Future research is needed to investigate the interrelations between self- and social image and how exactly such a form of self-deception works in different MWR settings.

(2) Can MRW also be conceptualized and measured as a continuous construct? If so, how would such a conceptualization look like, and what would it imply? For purposes of simplicity, we conceptualize MWR as a binary construct in MWR-T: situations either contain MWR, or they do not. However, from a perspective of external validity, it is more plausible to think of MWR as a continuous construct, with real-life situations offering more or less MWR to the agent. Also, the different operationalizations of MWR (see Appendix, Table A2) most likely induce different degrees of MWR. This may depend on several factors. For example, the different methods used to induce MWR seem to differ with regards to the observability of the agent's behavior, the effort needed to exploit MWR, and the possibility to fool oneself in addition to others about one's intentions (see previous discussion on the role of social versus self-image). For a more in-depth discussion of this idea, see Supplementary Material C (available online).

(3) Connected to the preceding point, one may question whether the proposed relationships are of linear form (as specified for Propositions 1–6) or whether they are better captured by more complex functions. For instance, it is conceivable that relationships, such as the one of MWR and social image concerns (Proposition 6), are better defined by threshold/hurdle models (as applied in research testing selfish behavior at different thresholds of personality traits, such as self-control and attention, or different beliefs about the costs of prosocial behavior; Martinsson et al., 2012; Moradi & Nesterov, 2017; Spiekermann & Weiss, 2016). Similarly, it needs to be tested whether other-regarding preferences indeed modulate the relevance of social image concerns with the proposed inverse U-shaped quadratic function (Proposition 7) or rather in a different manner.

4.2.2. Possibilities for extending and differentiating MWR-T

Once the described open questions have been answered, additional extensions and specifications of MWR-T can be attempted.

4.2.2.1. Possible extensions

Regarding the proposed boundary conditions of MWR-T, it would be interesting to test what happens if one relaxes the auxiliary assumptions that the action space does not include sharing above the equal split, taking options, or the possibility for moral balancing (i.e., repeated social decisions). Guiding questions here would be: Are more-than-equal shares just not socially demanded (Andreoni & Bernheim, 2009), or even socially undesired (Duncan, 2009; Tasimi et al., 2015)? Does including unethical options in the choice set induce feelings of entitlement (Cullis et al., 2012), because it changes the reference point for what counts as selfish (e.g., Bardsley, 2008)? Does this always reduce fair shares (Cappelen et al., 2013; List, 2007; Zhang & Ortmann, 2014)? Does repeated exposure to MWR result in moral balancing (Birkelund & Cherry, 2020)? Extensions of MWR-T that answer these questions would increase the theory's ecological validity and scope of applicability.

In addition, the main *mechanism of anticipated social image damage* from selfishness could be extended to include anticipated benefits to the agent's social image resulting from acting prosocially, because MWR could render one's social image generally less malleable. In other words, agents may not only give less under MWR, because they are less afraid of image damage, but also because they expect less benefit to their social image from being prosocial. This extension of MWR-T may be relevant for behavioral predictions (e.g., loss aversion; Tversky & Kahneman, 1991) and the role of individual differences (Sassenberg & Hansen, 2007).

4.2.2.2. Possible differentiations

Differentiations of MWR-T could concern the *mechanism of perceived social norms* by distinguishing between injunctive and descriptive norms (Jacobson et al., 2011; Reno et al., 1993; Smith et al., 2012). Currently, MWR-T only considers perceived injunctive norms, that is, an agent's perception of what others find socially appropriate. However, actions are also influenced by perceived descriptive norms, that is, perceptions of what others typically do in a situation (e.g., Bicchieri et al., 2022). Future research should tease apart how the two different types of norm perceptions relate to the effect of MWR.

We also see potential for a more fine-grained specification of the *emotional mechanism*, such as an elucidation of the role of *shame*. Similar to guilt, shame is a negative self-directed moral emotion (Tangney et al., 2007), and some researchers have in fact invoked shame as a partial explanation of the MWR effect (e.g., Bonner et al., 2017; Regner, 2021). In addition, other individual differences might come into play in modulating the effect of MWR. These could be differences in guilt proneness (Regner, 2021), HEXACO factors (Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience; e.g., Ashton et al., 2014), social norm espousal (Bizer et al., 2014), need for cognition (Petty et al., 2009), or (social) loss and reward processing (e.g., Boyce et al., 2016; BIS/BAS; Fricke & Vogel, 2020; Sassenberg & Hansen, 2007). Future theorizing should consider these distinctions and test empirically for their unique contributions.

4.3. How does MWR-T relate to other theories?

In addition to fully specifying a theory of interest, it is also important to clarify the relations of the specified theory to other theories in the field. While a comprehensive theory comparison is beyond the goals of this article, we will at least sketch the relations of MWR-T to a number of other relevant theories. Specifically, we show that MWR-T captures unique aspects of the moral decision-making process. For instance, it applies ideas about a combined impact of situational factors and individual differences known from moral judgment theory (e.g., Kohlberg et al., 1983) to the behavioral domain. It also adds to behavioral theory (e.g., Ajzen, 1991) as it goes beyond the interactive effect of attitudes and norm perceptions by defining situational circumstances which change norm perceptions, social image anticipations, and anticipatory emotional reactions.

Another theory closely related to MWR-T that therefore demands careful assessment is Bandura's (1986, 1991; see also Bandura et al., 1996) social-cognitive theory (SCT) of moral thought and action, which is an application of Bandura's earlier theory of social learning (Bandura, 1977) to moral self-regulation and disengagement. SCT proposes a self-regulatory system similar to the proposed mediating mechanisms of MWR-T. Specifically, Bandura proposes that people judge their own moral (mis)conduct by comparing their actions to internal and external reference standards that may include the other-regarding preferences and perceived social norms specified in MWR-T. In case of negative self-evaluation, people can react by, for example, self-sanctioning (corresponding to the feeling of guilt in MWR-T). As in MWR-T, these processes can be anticipatory and prevent misconduct. SCT also assumes that agents are 'morally flexible' and distinguishes between four different strategies of 'moral disengagement', which include, as in MWR-T, the strategies of ignoring harmful consequences and obscuring one's causal role in bringing about such consequences. Moreover, another strategy specified in SCT is that of 'moral justification', which could be understood as agents using an ends-justify-means logic where they portray the ends as something socially desirable (e.g., 'It is alright to fight to protect your friends'; Bandura et al., 1996, Appendix). While Bandura describes this as misconstruing the *action*, it could also be understood as construing an alternative (socially desirable) *intention* which would match MWR-T's proposition that MWR (as a situation obscuring the agent's intention) allows for the perception of changed social norms. Nevertheless, MWR-T differs in several important aspects from SCT: (1) Whereas SCT describes trait-like, stable strategies, that is, a 'proneness to moral disengagement' (Bandura et al., 1996, p. 367), and vaguely explains the possibility for moral flexibility with the moral complexity of many real-life situations, MWR-T highlights and explicitly specifies

https://doi.org/10.1017/jdm.2024.16 Published online by Cambridge University Press

the type of situational characteristics that are necessary for allowing people to be morally flexible in their behavior. (2) Whereas SCT focuses on reprehensible or transgressive behavior (stealing, lying, physical aggression), MWR-T explicitly restricts the action space to decisions ranging from maximum selfishness (but not malevolence) to egalitarian behavior. (3) Whereas diffusion of responsibility is a disengagement strategy in SCT, it is excluded as MWR operationalization. (4) Whereas SCT highlights the importance of *self*-image concerns, MWR-T emphasizes agents' desire to protect their *social* image. (5) Whereas MWR-T explicitly specifies the importance of individual differences in social image concerns and their interaction with other-regarding preferences as moderators of the effect of MWR on behavior, these propositions are not part of SCT. Apart from these differences in the content of the two theories, MWR-T could be regarded as having a higher level of specification than SCT, because it (6) disentangles input and output of the 'judgment' stage (i.e., social norm perceptions and anticipated social image damage, respectively), (7) offers precise concept definitions and clear operationalizations, and (8) specifies boundary conditions for the proposed effects. These differences show that MWR-T is not simply a disguised version of moral disengagement, but a theory of its own. However, given that the two theories agree on several core aspects, their integration into a single theory should be considered as a longer-range goal. Our theory specification can be seen as a first step in this direction. However, to fully identify the possibilities for theory integration, a careful proposition-based specification of SCT is needed as well.

The importance of situational justifications for immoral behavior stated by MWR-T has also been highlighted by attribution theory (Kelley, 1967, 1987). According to the so-called *discounting principle* proposed by Kelley (1987), observers tend to discount or minimize the role of a possible cause of an agent's behavior if other plausible causes are present, because they are unsure of the real cause. This suggests that obfuscating the link between the outcome and the selfish motive (i.e., MWR) in case of observable behavior actually works by making non-selfish causes of the behavior seem more plausible to observers. Indeed, people are judged less harshly by others when behaving selfishly under MWR (versus no MWR; Bartling et al., 2014). Furthermore, assuming that agents are (at least implicitly) aware of the discounting principle in observers, attribution theory can explain how agents infer reduced social image damage in situations with MWR.

MWR-T could also be incorporated into broader theoretical frameworks such as the framework of utility theory (e.g., Fishburn, 1970). The basic proposition of utility theory is that people choose actions that they believe maximize utility. Crucially, expected utilities resulting from a specific decision depend on agents' preferences and situational circumstances. Originally, utility only referred to self-serving gain. However, experimental evidence from behavioral economics and psychology broadened this concept: Observing a multitude of prosocial behaviors led researchers to argue against a purely selfishly motivated homo economicus and develop several social preference theories (Fehr & Fischbacher, 2002; Fehr & Schmidt, 1999; Murphy & Ackermann, 2014; Van Lange, 1999). These theories tried to explain prosociality by adding social preferences to the factors determining overall utility. Though there are different approaches to model how and why people have prosocial preferences, they all share the idea that agents who behave prosocially gain utility from behaving prosocially. However, research on MWR indicates that this does not explain all prosocial behavior. From existing research results, we have derived our propositions that situations without MWR increase the perceived social inappropriateness of selfish behavior as well as the anticipated social image damage and anticipatory guilt connected with selfish behavior, all of which carry their own disutility. According to MWR-T, the degree to which these disutilities (and their changes in the presence of MWR) enter into the overall utility of a specific action depends on the weight attached to them by an agent's social image concerns. Our theory specification also spells out how the role of social image concerns is moderated by other-regarding preferences.

In sum, although the mentioned alternative theories provide valuable insights into several aspects of moral judgment and decision-making, MWR-T is still needed to explain in detail the behavioral effect of MWR and its underlying psychological mechanisms. More generally speaking, theories could be formalized as sets of equations (e.g., Borsboom et al., 2021) or verbal propositions (e.g., West

https://doi.org/10.1017/jdm.2024.16 Published online by Cambridge University Press

et al., 2019). Although utility theories are often expressed in econometric equations, we decided to employ a verbal proposition-based theory specification for MWR-T. Among others, this type of theory specification has the advantage of reaching a broader readership, thus fostering interdisciplinary collaborations. We hope that our theory specification will serve as a blueprint for other theory specifications. The more theories follow verbal proposition-based specification rules, the easier it will be to connect them to related theories and integrate them into the network of theories. Furthermore, by relating our theory to several other potentially relevant theories, we were able to identify overlaps as well as unique contributions. Because the comparison of a specified theory to other relevant theories elucidates the connections between different theories and helps to spot potentials for theory integration, we propose that it should become a standard component of every theory specification. As a discipline, and to heighten efficiency, we should strive toward more unified theoretical frameworks functioning as catalysts of knowledge about human behavior and its underlying psychological mechanisms.

5. Conclusion

Like many theories, the original formulation and explanation of the MWR effect (DWK) suffers from underspecification. In the present article, we argue that the method of verbal proposition-based theory specification (Glöckner & Betsch, 2011) is helpful to remediate this underspecification while maintaining accessibility to a broad readership. Our specification of DWK's original formulation reveals several problems and loopholes pertaining to all aspects of a well-specified theory. We assume that this may explain the vast heterogeneity in study designs testing the MWR effect and we argue that this hampers the interpretability and comparability of study results and, ultimately, scientific progress. To allow for a common understanding of MWR, its behavioral effects, and psychological explanations, and to enable stringent empirical tests, we proposed a fully specified MWR-T. This also allowed us to identify open questions and derive an empirical roadmap. For instance, future research should critically assess the role of self- versus social image concerns in MWR-T, the relevance of the action space available to agents, and the exact type of perceived social norms and moral emotions driving MWR effects. We also compared MWR-T to related theories in the social sciences and concluded that MWR-T captures unique and important aspects of social decision-making in situations providing MWR. However, we suggest that future work could integrate MWR-T with the SCT of moral thought and action (Bandura, 1986, 1991; Bandura et al., 1996). With these contributions, we hope to stimulate fruitful and efficient future research on MWR. In addition, we hope that the present article demonstrates the usefulness of the verbal proposition-based specification of theories, which may motivate other researchers to use the same method.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/jdm.2024.16.

Acknowledgements. We would like to thank three anonymous reviewers for their insightful comments and suggestions which have contributed to the improvement of this work. We truly appreciate their time, effort, and expertise in reviewing our manuscript.

Funding statement. This research received no specific grant funding from any funding agency, commercial or not-for-profit sectors.

Competing interest. The authors declare none.

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T.

Andreoni, J., & Bernheim, B. D. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, *77*(5), 1607–1636. https://doi.org/10.3982/ECTA7384.

Andreoni, J., & Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, *70*(2), 737–753.

Asendorpf, J. B., Conner, M., de Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., van Aken, M. A. G., Weber, H., & Wicherts, J. M. (2016). *Recommendations for increasing replicability in psychology*. American Psychological Association. https://doi.org/10.1037/14805-038.

Ashton, M. C., Lee, K., & de Vries, R. E. (2014). The HEXACO honesty-humility, agreeableness, and emotionality factors: A review of research and theory. *Personality and Social Psychology Review*, *18*(2), 139–152. https://doi.org/10.1177/1088868314523838.

Balzer, W., & Moulines, C. U. (1996). *Structuralist theory of science: Focal issues, new results*. Walter de Gruyter.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191.

Bandura, A. (1986). *Social foundations of thought and action* (pp. 23–28). Englewood Cliffs.

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes*, *50*, 248–287.

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Mechanisms of moral disengagement in the exercise of moral agency. *Journal of Personality and Social Psychology*, *71*(2), 364.

Bardsley, N. (2008). Dictator game giving: Altruism or artefact? *Experimental Economics*, *11*(2), 122–133. https://doi.org/10.1007/s10683-007-9172-2.

Bartling, B., Engl, F., & Weber, R. A. (2014). Does willful ignorance deflect punishment? An experimental study. *European Economic Review*, *70*, 512–524. https://doi.org/10.1016/j.euroecorev.2014.06.016.

Bell, E., Norwood, F. B., & Lusk, J. L. (2017). Are consumers willfully ignorant about animal welfare? *Animal Welfare*, *26*(4), 399–402. https://doi.org/10.7120/09627286.26.4.399.

Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, *96*(5), 1652–1678. https://doi.org/10.1257/aer.96.5.1652.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C., Dimant, E., Gächter, S., & Nosenzo, D. (2022). Social proximity and the erosion of norm compliance. *Games and Economic Behavior*, *132*, 59–72. https://doi.org/10.1016/j.geb.2021.11.012.

Birkelund, J., & Cherry, T. L. (2020). Institutional inequality and individual preferences for honesty and generosity. *Journal of Economic Behavior & Organization*, *170*, 355–361. https://doi.org/10.1016/j.jebo.2019.12.014.

Bizer, G. Y., Magin, R. A., & Levine, M. R. (2014). The Social-Norm Espousal Scale. *Personality and Individual Differences*, *58*, 106–111. https://doi.org/10.1016/j.paid.2013.10.014.

Bodner, R., & Prelec, D. (2003). Self-signaling and diagnostic utility in everyday decision making. In *The psychology of economic decisions: Rationality and well-being* (Vol. *1*, pp. 105–126). Oxford University Press.

Bolton, G. E., Kusterer, D. J., & Mans, J. (2019). Inflated reputations: Uncertainty, leniency, and moral wiggle room in trader feedback systems. *Management Science*, *65*(11), 5371–5391. https://doi.org/10.1287/mnsc.2018.3191.

Bonner, J. M., Greenbaum, R. L., & Quade, M. J. (2017). Employee unethical behavior to shame as an indicator of self-image threat and exemplification as a form of self-image protection: The exacerbating role of supervisor bottom-line mentality. *Journal of Applied Psychology*, *102*(8), 1203–1221. https://doi.org/10.1037/apl0000222.

Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*(4), 756–766. https://doi.org/10.1177/1745691620969647.

Boyce, C. J., Wood, A. M., & Ferguson, E. (2016). Individual differences in loss aversion: Conscientiousness predicts how life satisfaction responds to losses versus gains in income. *Personality and Social Psychology Bulletin*, *42*(4), 471–484. https://doi.org/10.1177/0146167216634060.

Bursztyn, L., & Jensen, R. (2017). Social image and economic behavior in the field: Identifying, understanding, and shaping social pressure. *Annual Review of Economics*, *9*(1), 131–153. https://doi.org/10.1146/annurev-economics-063016-103625.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S. & Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z.

Cappelen, A. W., Nielsen, U. H., Sørensen, E. Ø., Tungodden, B., & Tyran, J.-R. (2013). Give and take in dictator games. *Economics Letters*, *118*(2), 280–283. https://doi.org/10.1016/j.econlet.2012.10.030.

Cerrone, C., & Engel, C. (2019). Deciding on behalf of others does not mitigate selfishness: An experiment. *Economics Letters*, *183*, 108616. https://doi.org/10.1016/j.econlet.2019.108616.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, *117*(3), 817–869. https://doi.org/10.1162/003355302760193904.

Conrads, J., & Irlenbusch, B. (2013). Strategic ignorance in ultimatum bargaining. *Journal of Economic Behavior & Organization*, *92*, 104–115. https://doi.org/10.1016/j.jebo.2013.05.010.

Cullis, J., Jones, P., & Savoia, A. (2012). Social norms and tax compliance: Framing the decision to pay tax. *The Journal of Socio-Economics*, *41*(2), 159–168. https://doi.org/10.1016/j.socec.2011.12.003.

https://doi.org/10.1017/jdm.2024.16 Published online by Cambridge University Press

D'Adda, G., Gao, Y., Golman, R., & Tavoni, M. (2018). It's so hot in here: Information avoidance, moral wiggle room, and high air conditioning usage (Working Paper ID 3149330). Social Science Research Network. https://papers.ssrn.com/abstract=3149330.

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*(1), 67–80. https://doi.org/10.1007/s00199-006-0153-z.

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, *8*(4), 377–383. https://doi.org/10.1037/h0025589.

Duncan, B. (2009). Secret santa reveals the secret side of giving. *Economic Inquiry*, *47*(1), 165–181. https://doi.org/10.1111/j.1465-7295.2008.00145.x.

Ehrich, K. R., & Irwin, J. R. (2005). Willful ignorance in the request for product attribute information. *Journal of Marketing Research*, *42*(3), 266–277. https://doi.org/10.1509/jmkr.2005.42.3.266.

Exley, C. L. (2016). Excusing selfishness in charitable giving: The role of risk. *The Review of Economic Studies*, *83*(2), 587–628. https://doi.org/10.1093/restud/rdv051.

Fehr, E., & Fischbacher, U. (2002). Why social preferences matter – The impact of non-selfish motives on competition, cooperation and incentives. *The Economic Journal*, *112*(478), 1–33. https://doi.org/10.1111/1468-0297.00027.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868. https://doi.org/10.1162/003355399556151.

Feiler, L. (2014). Testing models of information avoidance with binary choice dictator games. *Journal of Economic Psychology*, *45*, 253–267. https://doi.org/10.1016/j.joep.2014.10.003.

Fishburn, P. C. (1970). *Utility theory for decision making*. Research Analysis Corp McLean VA.

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, *6*(3), 347–369. https://doi.org/10.1006/game.1994.1021.

Fricke, K., & Vogel, S. (2020). How interindividual differences shape approach-avoidance behavior: Relating self-report and diagnostic measures of interindividual differences to behavioral measurements of approach and avoidance. *Neuroscience & Biobehavioral Reviews*, *111*, 30–56. https://doi.org/10.1016/j.neubiorev.2020.01.008.

Friedrichsen, J., & Engelmann, D. (2013). Who cares for social image? Interactions between intrinsic motivation and social image concerns (SSRN Scholarly Paper 2371250). https://doi.org/10.2139/ssrn.2371250.

Garcia, T., Massoni, S., & Villeval, M. C. (2020). Ambiguity and excuse-driven behavior in charitable giving. *European Economic Review*, *124*, 103412. https://doi.org/10.1016/j.euroecorev.2020.103412.

Gärtner, M., & Sandberg, A. (2017). Is there an omission effect in prosocial behavior? A laboratory experiment on passive vs. active generosity. *PLOS ONE*, *12*(3), 1–21. https://doi.org/10.1371/journal.pone.0172496.

Glöckner, A., & Betsch, T. (2011). The empirical content of theories in judgment and decision making: Shortcomings and remedies. *Judgment and Decision Making*, *6*(8), 711–721.

Gollwitzer, M., & Schwabe, J. (2020). Context dependency as a predictor of replicability [Working Paper]. https://doi.org/10.31234/osf.io/53yhg.

Gotowiec, S., & van Mastrigt, S. (2019). Having versus doing: The roles of moral identity internalization and symbolization for prosocial behaviors. *The Journal of Social Psychology*, *159*(1), 75–91. https://doi.org/10.1080/00224545.2018.1454394.

Gross, J., & Vostroknutov, A. (2022). Why do people follow social norms? *Current Opinion in Psychology*, *44*, 1–6. https://doi.org/10.1016/j.copsyc.2021.08.016.

Grossman, Z. (2010). Strategic ignorance and the robustness of social preferences. https://escholarship.org/uc/item/60b93868.

Grossman, Z. (2014). Strategic ignorance and the robustness of social preferences. *Management Science*, *60*(11), 2659–2665. https://doi.org/10.1287/mnsc.2014.1989.

Grossman, Z. (2015). Self-signaling and social-signaling in giving. *Journal of Economic Behavior & Organization*, *117*, 26–39. https://doi.org/10.1016/j.jebo.2015.05.008.

Grossman, Z., & van der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association*, *15*(1), 173–217. https://doi.org/10.1093/jeea/jvw001.

Haley, K. J., & Fessler, D. M. T. (2005). Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*, *26*(3), 245–256. https://doi.org/10.1016/j.evolhumbehav.2005.01.002.

Hauge, K. E. (2016). Generosity and guilt: The role of beliefs and moral standards of others. *Journal of Economic Psychology*, *54*, 35–43. https://doi.org/10.1016/j.joep.2016.03.001.

Jacobson, R. P., Mortensen, C. R., & Cialdini, R. B. (2011). Bodies obliged and unbound: Differentiated response tendencies for injunctive and descriptive social norms. *Journal of Personality and Social Psychology*, *100*(3), 433–448. https://doi.org/10.1037/a0021470.

Jordan, J. J., & Rand, D. G. (2020). Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *Journal of Personality and Social Psychology*, *118*, 57–88. https://doi.org/10.1037/pspi0000186.

Kelley, H. H. (1967). Attribution theory in social psychology. *Nebraska Symposium on Motivation*, *15*, 192–238.

Kelley, H. H. (1987). Attribution in social interaction. In *Attribution: Perceiving the causes of behavior* (pp. 1–26). Lawrence Erlbaum Associates, Inc.

Kohlberg, L., Levine, C., & Hewer, A. (1983). *Moral stages: A current formulation and a response to critics*. S.Karger AG. https://karger.com/books/book/3475/Moral-StagesA-Current-Formulation-and-a-Response.

https://doi.org/10.1017/jdm.2024.16 Published online by Cambridge University Press

Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, *11*(3), 495–524. https://doi.org/10.1111/jeea.12006.

Krysowski, E., & Tremewan, J. (2021). Why does anonymity make us misbehave: Different norms or less compliance? *Economic Inquiry*, *59*(2), 776–789. https://doi.org/10.1111/ecin.12955.

Larson, T., & Capra, C. M. (2009). Exploiting moral wiggle room: Illusory preference for fairness? A comment. *Judgment and Decision Making*, *4*(6), 467–474.

Lazear, E. P., Malmendier, U., & Weber, R. A. (2012). Sorting in experiments with application to social preferences. *American Economic Journal: Applied Economics*, *4*(1), 136–163. https://doi.org/10.1257/app.4.1.136.

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, *115*(3), 482–493. https://doi.org/10.1086/519249.

Martinsson, P., Myrseth, K. O. R., & Wollbrant, C. (2012). Reconciling pro-social vs. selfish behavior: On the role of self-control. *Judgment and Decision Making*, *7*(3), 304–315.

Matthey, A., & Regner, T. (2011). Do I really want to know? A cognitive dissonance-based explanation of other-regarding behavior. *Games*, *2*(1), 114–135. https://doi.org/10.3390/g2010114.

Matthey, A., & Regner, T. (2014). More than outcomes: The role of self-image in other-regarding behavior (Working Paper 2014–036). Jena Economic Research Papers. https://www.econstor.eu/handle/10419/108543.

Momsen, K., & Ohndorf, M. (2020). When do people exploit moral wiggle room? An experimental analysis of information avoidance in a market setup. *Ecological Economics*, *169*, 106479. https://doi.org/10.1016/j.ecolecon.2019.106479.

Moradi, H., & Nesterov, A. (2017). Moral wiggle room reverted: Information avoidance is myopic (SSRN Scholarly Paper ID 3168630). Social Science Research Network. https://papers.ssrn.com/abstract=3168630.

Müller, S., & Moshagen, M. (2019). True virtue, self-presentation, or both? A behavioral test of impression management and overclaiming. *Psychological Assessment*, *31*(2), 181–191. https://doi.org/10.1037/pas0000657.

Murphy, R. O., & Ackermann, K. A. (2014). Social value orientation: Theoretical and measurement issues in the study of social preferences. *Personality and Social Psychology Review*, *18*(1), 13–41. https://doi.org/10.1177/1088868313501745.

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. J. J. (2011). Measuring social value orientation. *Judgment and Decision Making*, *6*(8), 771–781.

Ockenfels, A., & Werner, P. (2012). 'Hiding behind a small cake' in a newspaper dictator game. *Journal of Economic Behavior & Organization*, *82*(1), 82–85. https://doi.org/10.1016/j.jebo.2011.12.008.

Petty, R. E., Brinol, P., Loersch, C., & McCaslin, M. J. (2009). The need for cognition. In *Handbook of individual differences in social behavior* (pp. 318–329). The Guilford Press.

Popper, K. (1934). *Logik der Forschung. Zur Erkenntnistheorie der modernen Naturwissenschaft*. Verlag von Julius Springer.

Regner, T. (2021). What's behind image? Toward a better understanding of image-driven behavior. *Frontiers in Psychology*, *12*, 614575. https://doi.org/10.3389/fpsyg.2021.614575.

Reno, R. R., Cialdini, R. B., & Kallgren, C. A. (1993). The transsituational influence of social norms. *Journal of Personality and Social Psychology*, *64*(1), 104–112. https://doi.org/10.1037/0022-3514.64.1.104.

Sassenberg, K., & Hansen, N. (2007). The impact of regulatory focus on affective responses to social discrimination. *European Journal of Social Psychology*, *37*(3), 421–444. https://doi.org/10.1002/ejsp.358.

Smaldino, P. (2019). Better methods can't make up for mediocre theory. *Nature*, *575*(7781), 9. https://doi.org/10.1038/d41586-019-03350-5.

Smith, J. R., Louis, W. R., Terry, D. J., Greenaway, K. H., Clarke, M. R., & Cheng, X. (2012). Congruent or conflicted? The impact of injunctive and descriptive norms on environmental intentions. *Journal of Environmental Psychology*, *32*(4), 353–361. https://doi.org/10.1016/j.jenvp.2012.06.001.

Spiekermann, K., & Weiss, A. (2016). Objective and subjective compliance: A norm-based explanation of 'moral wiggle room.'. *Games and Economic Behavior*, *96*, 170–183. https://doi.org/10.1016/j.geb.2015.11.007.

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology*, *58*(1), 345–372. https://doi.org/10.1146/annurev.psych.56.091103.070145.

Tankard, M. E., & Paluck, E. L. (2016). Norm perception as a vehicle for social change. *Social Issues and Policy Review*, *1*(10), 181–211. https://doi.org/10.1111/sipr.12022.

Tasimi, A., Dominguez, A., & Wynn, K. (2015). Do-gooder derogation in children: The social costs of generosity. *Frontiers in Psychology*, *6*, 1036. https://doi.org/10.3389/fpsyg.2015.01036.

tho Pesch, F., & Dana, J. (2024). Attributional ambiguity reduces charitable giving by relaxing social norms. *Journal of Experimental Social Psychology*, *110*, 104530.

Thunström, L., Veld, K. v. 't., Shogren, J. F., & Nordström, J. (2014). On strategic ignorance of environmental harm and social norms. *Revue d'economie Politique*, *124*(2), 195–214. https://doi.org/10.3917/redp.242.0195.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, *106*(4), 1039–1061. https://doi.org/10.2307/2937956.

van der Weele, J. J., Kulisa, J., Kosfeld, M., & Friebel, G. (2014). Resisting moral wiggle room: How robust is reciprocal behavior? *American Economic Journal: Microeconomics*, *6*(3), 256–264. https://doi.org/10.1257/mic.6.3.256.

Van Lange, P. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology*, *77*(2), 337–349. https://doi.org/10.1037/0022-3514.77.2.337.

https://doi.org/10.1017/jdm.2024.16 Published online by Cambridge University Press

West, R., Godinho, C. A., Bohlen, L. C., Carey, R. N., Hastings, J., Lefevre, C. E., & Michie, S. (2019). Development of a formal system for representing behaviour-change theories. *Nature Human Behaviour*, *3*(5), 526–536. https://doi.org/10.1038/s41562-019-0561-2.

Winterich, K. P., Aquino, K., Mittal, V., & Swartz, R. (2013). When moral identity symbolization motivates prosocial behavior: The role of recognition and moral identity internalization. *Journal of Applied Psychology*, *98*(5), 759–770. https://doi.org/10.1037/a0033177.

Zhang, L., & Ortmann, A. (2014). The effects of the take-option in dictator-game experiments: A comment on Engel's (2011) meta-study. *Experimental Economics*, *17*(3), 414–420. https://doi.org/10.1007/s10683-013-9375-7.

A. Appendix

Table A1. *Propositions of MWR-T.*

Proposition	Antecedence	Consequence
For all agents and situations specified in the respective auxiliary assumptions:		
Behavioral proposition: Effect of MWR		
1	IF MWR (instead of no MWR)	THEN increase in selfish behavior between these situations
Image mechanism: Reduction in anticipated social image damage		
2a	IF MWR (instead of no MWR)	THEN decrease in anticipated social image damage attached to selfish behavior between these situations
2b	IF decrease in anticipated social image damage attached to selfish behavior between situations	THEN increase in selfish behavior between these situations
Normative mechanism: Change in perceived social norms		
3a	IF MWR (instead of no MWR)	THEN change in perceived social norms (decrease in the perception of selfish behavior as socially inappropriate) between these situations
3b	IF change in perceived social norms (decrease in the perception of selfish behavior as socially inappropriate) between situations	THEN increase in selfish behavior between these situations
Emotional mechanism: Reduction in anticipatory guilt		
4a	IF MWR (instead of no MWR)	THEN decrease in anticipatory guilt between these situations
4b	IF decrease in anticipatory guilt between situations	THEN increase in selfish behavior between these situations

Continued

Table A1. *Continued.*

Proposition	Antecedence	Consequence
Psychological mechanism interrelations		
5a	IF change in perceived social norms (decrease in the perception of selfish behavior as socially inappropriate) between situations	THEN decrease in anticipated social image damage attached to selfish behavior between these situations
5b	IF decrease in anticipated social image damage attached to selfish behavior between situations	THEN decrease in anticipatory guilt between these situations
Individual differences		
6	THE HIGHER the social image concerns	THE GREATER the behavioral effect of MWR
7	THE MORE extreme (i.e., selfish or prosocial) the other–regarding preferences	THE SMALLER the relevance of social image concerns for the behavioral effect of MWR
		AND THE SMALLER the behavioral effect of MWR

Note: The formulations in this table are abbreviations of full sentences. For example, Proposition 1 reads *“If a person is in a situation containing moral wiggle room, then this person will show more selfish behavior relative to a situation not containing MWR.”* Similarly, Proposition 6 reads *“The higher a person’s (dispositional) social image concerns, the greater will be the effect of that person being in a situation containing MWR compared to a situation not containing MWR on the person’s behavior.”* All proposed links should be viewed as probabilistic rather than deterministic. Furthermore, for simplicity, the relationships in Propositions 1–6 are conceptualized to be of linear form, while Proposition 7 explicitly specifies an inverse-u-shaped relationship.

https://doi.org/10.1017/jdm.2024.16 Published online by Cambridge University Press

Table A2. *Concept definitions and operationalizations (per proposition of MWR-T).*

Proposition	Concept label	Verbal definition	Measurement/operationalization[1]

1,
2a,
3a,
4a,
6,
7
	MWR (Moral Wiggle Room)	Situational characteristics that obfuscate the signal which the outcome of an own-payoff-maximizing (i.e., potentially selfish) behavior sends to others about one's intention to behave selfishly; consequently, no MWR describes situations with full transparency between intention and outcome (i.e., the baseline setting)	Various options: Outcome ignorance (Dana et al., 2007; Vu et al., 2023);
			Delegation (Erat, 2013; Hamman et al., 2010);
			Default implementation (Gärtner & Sandberg, 2017);
			Exiting (Andreoni et al., 2017; Dana et al., 2006; DellaVigna et al., 2012);
			Omission (Gärtner & Sandberg, 2017);
			Decision implementation uncertainty (Matthey & Regner, 2015);
			Additional choice attributes (Exley, 2016; Haisley & Weber, 2010; Snyder et al., 1979; tho Pesch & Dana, 2024);
			Information asymmetry (Güth & Huck, 1997; Ockenfels & Werner, 2012);
			Hybrid treatment (Dana et al., 2007; Gärtner & Sandberg, 2017)

1, 2a, 2b, 3a, 3b, 4b, 5a, 5b	Selfish behavior	Choosing a (more) selfish distribution option over one or several (more) prosocial distribution option(s); a (more) selfish option always yields a higher payoff for the agent and a lower payoff for the recipient, whereas (more) prosocial options are less profitable for the agent but reduce the outcome inequality between the agent and recipient; an increase in selfish behavior is estimated either as an increase in the likelihood of selfish behavior or the average degree of selfish behavior	Three options: (1) Binary decision (one monetary distribution option yielding a higher payoff to the agent but a lower payoff for a passive recipient, compared to another, more equitable distribution option); (2) Ordinal variable presenting several options; (3) Continuous variable allowing for a free decision

Continued

https://doi.org/10.1017/jdm.2024.16 Published online by Cambridge University Press

Table A2. *Continued.*

Proposition	Concept label	Verbal definition	Measurement/operationalization[1]
2a, 2b, 5a, 5b	Anticipated social image damage	Agents' expectations of how negatively others will judge them in terms of their morality due to their (selfish) decisions; presumably generated by a cognitive appraisal process taking place when agents consider the different distribution options before actually deciding	"How much do you think the image others hold of you will change if you choose option A in this scenario?" (Scale: -7 = others would think much more negatively of me, 0 = it would not change, 7 = others would think much more positively of me); procedure: randomized, counterbalanced (in multiple scenarios with and without MWR and with the choice of option A constituting selfish or prosocial behavior); for the avoidance of response bias (in line with decision): third-party responses OR first-party self-report items for hypothetical scenarios (vignettes) that are temporally separated from the measurement of behavior
3a, 3b, 5a	Perceived social norms	Perception of behavioral rules based on social consensus regarding appropriate behavior (injunctive norm)	Two options: (1) For the avoidance of response bias (in line with decision): third parties' normative perceptions of what most people rated as appropriate (perceived injunctive norm) for the respective decision in the respective setting OR first-party self-report normative perceptions for hypothetical scenarios (vignettes) that are temporally separated from the measurement of behavior– using a norm-elicitation method is advised in either case (Krupka & Weber, 2013); (2) Asking third parties to indicate their willingness to punish the respective decision in the respective setting (Carpenter & Matthews, 2009; Fehr & Fischbacher, 2004; House et al., 2020)–more indirect measure, but holds the advantage of informing about how strongly the norm is enforced in each setting

Continued

Table A2. *Continued.*

Proposition	Concept label	Verbal definition	Measurement/operationalization[1]
4a, 4b, 5b	Anticipatory guilt[2]	Negative emotion, experienced when contemplating decision options that violate perceived moral standards; in situations, in which the relevant decision pits own-payoff-maximization against fairness considerations, anticipatory guilt is expected to arise when the agent contemplates acting selfishly	Three options: (1) State Shame and Guilt Scale (Marschall et al., 1994; note the need for modification to measure anticipatory guilt)-assessment in third parties for each setting (with vs. without MWR); (2) Activated Displeasure[3] scale of the 12–point circumplex structure of core affect (12-PAC; Yik et al., 2011) in the “Describes Me” format– assessment in third parties for each setting (with vs. without MWR); (3) Activated Displeasure scale of the 12-point circumplex structure of core affect (12–PAC; Yik et al., 2011) in the “Agree” format-assessment in first parties (because no explicit reference to guilt that could be biased in first parties) for each setting (with vs. without MWR)
6, 7	Social image concerns	How much an agent cares about being evaluated positively by others; in a specific situation, an agent’s social image concerns are determined by the person’s dispositional social image concerns (i.e., how much the person generally cares about being positively evaluated) and by domain-specific factors (e.g., whether the relevant others are their own family vs. anonymous others)	Three options: (1) Conventional social desirability scales (e.g., the Social Desirability Scale by Stöber, 2001, or the Impression Management scale of the BIDR, Paulhus & Reid, 1991; but see Lanz et al., 2021); (2) Symbolization scale of the Moral Identity measure (Aquino & Reed, 2002); (3) Agent’s willingness to pay to remain anonymous to a third-party observer in case the observable outcome of their unobservable decision is unfair (i.e., profitable for themselves, but unprofitable for the passive recipient; Henry & Sonntag, 2015)

Continued

Table A2. *Continued.*

Proposition	Concept label	Verbal definition	Measurement/operationalization[1]
7	Other-regarding preferences	Agent's true preferences over the distribution of resources between themselves and a recipient; presumably independent of social image concerns and not influenced by MWR; in a specific situation, an agent's other-regarding preferences are determined by their dispositional other-regarding preferences (i.e., how much they generally care about themselves vs. others) as well as by domain-specific factors (e.g., whether the recipients are children, animals, or the environment; or whether the shared good is money, time, etc.)	To avoid confounding with image concerns[4]: implicit tests of prosociality (e.g., the Self vs. Other Interest Implicit Association Test, Thornton & Aknin, 2020)[5]

Note: Effects of MWR are assumed to exist at the individual level, but all links between the listed concepts are viewed as probabilistic (allowing for systematic and unsystematic variance). Therefore, all effects on the listed concepts are estimated at the group level.

[1] For short descriptions of all operationalizations of MWR, see Supplementary Material A, Table SA1 (available online).

[2] To avoid response bias in the assessment of such morally charged emotions, we suggest using a combination of different measures.

[3] Note that responses in line with the Activated Displeasure dimension also have been connected to self-reported behavioral inhibition (Yik et al., 2011).

[4] Typically, other-regarding preferences are measured via the conventional ring-measure of social value orientation (Murphy et al., 2011). However, this measure is very similar to dictator game decisions without MWR, meaning it confounds true other-regarding preferences with image concerns. Similar issues may arise when using other self-report measures that explicitly ask about other-regarding preferences, such as the altruism module proposed by Falk et al. (2016), the dispositional greed scale by Seuntjens et al. (2015), or the 16-item Prosocialness Scale by Caprara et al. (2005).

[5] Such measures seem to correlate well with everyday prosocial behavior that involves proactive involvement (e.g., registering oneself as blood donor), but less with donation decisions (presumably because image concerns play a role for donations, but are not confounded in the IAT).

https://doi.org/10.1017/jdm.2024.16 Published online by Cambridge University Press

Table A3. *Auxiliary assumptions (per proposition of MWR-T).*

Proposition	Auxiliary assumption
1–7	(1) The decision must have consequences for oneself and others, and the interests of these parties must conflict.
1, 2a, 3a, 4a, 6, 7	(2) The presence of MWR in the social decision situation must not restrict the agent's choice (i.e., their ability to implement any of the outcomes available without MWR).
2a–7	(3) The interaction between the agent and the recipient is nonstrategic and unilateral.
2a–7	(4) The behavioral space should range from maximally selfish distribution options (i.e., keeping the whole initial endowment) to egalitarian distribution options (i.e., 50:50 split).
1–7	(5) There are two independent and sometimes conflicting motives active in agents in the population: (a) an agent's preferences over payoff distributions (other-regarding preferences) and (b) an agent's social image concerns. Additional motives possibly influencing agents in a social decision situation must be controlled for or held constant.

Note: Auxiliary Assumptions 1-3 are the same as for the specification of DWK's original formulations (see Supplementary Material, Table SB3). Auxiliary Assumption 5 has been modified compared to the specification of DWK's original formulations. Auxiliary Assumption 4 has been added.

Reasons for the auxiliary assumptions: If Auxiliary Assumption 1 is not met, selfish behavior is indistinguishable from prosocial behavior. If Auxiliary Assumption 2 is not met, selfish behavior may be an artifact of not being able to choose freely. If Auxiliary Assumptions 3 - 5 are not met, behavior may be explained by additional confounding factors (e.g., expectation of reciprocity, other norms) instead of or in addition to the proposed mechanisms. Moreover, if Auxiliary Assumption 5 is not met, the behavioral effect of MWR (and its underlying mechanisms) may not be observable.

References (Appendix)

Andreoni, J., Rao, J. M., & Trachtman, H. (2017). Avoiding the ask: A field experiment on altruism, empathy, and charitable giving. *Journal of Political Economy*, *125*(3), 625–653. https://doi.org/10.1086/691703.

Aquino, K., & Reed, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*(6), 1423–1440. https://doi.org/10.1037//0022-3514.83.6.1423.

Caprara, G. V., Steca, P., Zelli, A., & Capanna, C. (2005). A new scale for measuring adults' prosocialness. *European Journal of Psychological Assessment*, *21*(2), 77–89. https://doi.org/10.1027/1015-5759.21.2.77.

Carpenter, J., & Matthews, P. H. (2009). What norms trigger punishment? *Experimental Economics*, *12*(3), 272–288. https://doi.org/10.1007/s10683-009-9214-z.

Dana, J., Cain, D. M., & Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes*, *100*(2), 193–201. https://doi.org/10.1016/j.obhdp.2005.10.001.

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*(1), 67–80. https://doi.org/10.1007/s00199-006-0153-z.

DellaVigna, S., List, J. A., & Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *The Quarterly Journal of Economics*, *127*(1), 1–56. https://doi.org/10.1093/qje/qjr050.

Erat, S. (2013). Avoiding lying: The case of delegated deception. *Journal of Economic Behavior & Organization*, *93*, 273–278. https://doi.org/10.1016/j.jebo.2013.03.035.

Exley, C. L. (2016). Excusing selfishness in charitable giving: The role of risk. *The Review of Economic Studies*, *83*(2), 587–628. https://doi.org/10.1093/restud/rdv051.

Falk, A., Becker, A., Dohmen, T. J., Huffman, D., & Sunde, U. (2016). The Preference Survey Module: A validated instrument for measuring risk, time, and social preferences (Working Paper ID 2725874). Social Science Research Network. https://papers.ssrn.com/abstract=2725874.

https://doi.org/10.1017/jdm.2024.16 Published online by Cambridge University Press

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87. https://doi.org/10.1016/S1090-5138(04)00005-4.

Gärtner, M., & Sandberg, A. (2017). Is there an omission effect in prosocial behavior? A laboratory experiment on passive vs. active generosity. *PLOS ONE*, *12*(3), 1–21. https://doi.org/10.1371/journal.pone.0172496.

Güth, W., & Huck, S. (1997). From ultimatum bargaining to dictatorship—An experimental study of four games varying in veto power. *Metroeconomica*, *48*(3), 262–299. https://doi.org/10.1111/1467-999X.00033.

Haisley, E. C., & Weber, R. A. (2010). Self-serving interpretations of ambiguity in other-regarding behavior. *Games and Economic Behavior*, *68*(2), 614–625. https://doi.org/10.1016/j.geb.2009.08.002.

Hamman, J. R., Loewenstein, G., & Weber, R. A. (2010). Self-interest through delegation: An additional rationale for the principal-agent relationship. *American Economic Review*, *100*(4), 1826–1846. https://doi.org/10.1257/aer.100.4.1826.

Henry, E., & Sonntag, J. (2015). *Measuring image concerns* (Working Paper ID 2663429). Social Science Research *Network.* https://papers.ssrn.com/abstract=2663429.

House, B. R., Kanngiesser, P., Barrett, H. C., Yilmaz, S., Smith, A. M., Sebastian-Enesco, C., Erut, A., & Silk, J. B. (2020). Social norms and cultural diversity in the development of third-party punishment. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1925), 20192794. https://doi.org/10.1098/rspb.2019.2794.

Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, *11*(3), 495–524. https://doi.org/10.1111/jeea.12006.

Lanz, L., Thielmann, I., & Gerpott, F. H. (2021). Are social desirability scales desirable? A meta-analytic test of the validity of social desirability scales in the context of prosocial behavior. *Journal of Personality*, *1–19*. https://doi.org/10.1111/jopy.12662.

Marschall, D. E., Saftner, J., & Tangney, J. P. (1994). *The State Guilt and Shame Scale* [Working Paper]. George Mason University.

Matthey, A., & Regner, T. (2015). More than outcomes: The role of self-image in other-regarding behavior. *Review of Behavioral Economics*, *2*(4), 353–378. https://doi.org/10.1561/105.00000038.

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. J. J. (2011). Measuring Social Value Orientation. *Judgment and Decision Making*, *6*(8), 771–781.

Ockenfels, A., & Werner, P. (2012). 'Hiding behind a small cake' in a newspaper dictator game. *Journal of Economic Behavior & Organization*, *82*(1), 82–85. https://doi.org/10.1016/j.jebo.2011.12.008.

Paulhus, D. L., & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology*, *60*(2), 307–317. https://doi.org/10.1037/0022-3514.60.2.307.

Seuntjens, T. G., Zeelenberg, M., van de Ven, N., & Breugelmans, S. M. (2015). Dispositional greed. *Journal of Personality and Social Psychology*, *108*(6), 917–933. https://doi.org/10.1037/pspp0000031.

Snyder, M. L., Kleck, R. E., Strenta, A., & Mentzer, S. J. (1979). Avoidance of the handicapped: An attributional ambiguity analysis. *Journal of Personality and Social Psychology*, *37*(12), 2297–2306. https://doi.org/10.1037/0022-3514.37.12.2297.

Stöber, J. (2001). The Social Desirability Scale-17 (SDS-17): Convergent validity, discriminant validity, and relationship with age. *European Journal of Psychological Assessment*, *17*(3), 222–232. https://doi.org/10.1027/1015-5759.17.3.222.

tho Pesch, F., & Dana, J. (2024). Attributional ambiguity reduces charitable giving by relaxing social norms. *Journal of Experimental Social Psychology*, *110*, 104530.

Thornton, E. M., & Aknin, L. B. (2020). Assessing the validity of the self versus other interest implicit association test. *PLOS ONE*, *15*(6), e0234032. https://doi.org/10.1371/journal.pone.0234032.

Vu, L., Soraperra, I., Leib, M., Van Der Weele, J., & Shalvi, S. (2023). Ignorance by choice: A meta-analytic review of the underlying motives of willful ignorance and its consequences. *Psychological Bulletin*, *149*(9–10), 611–635. https://doi.org/10.1037/bul0000398.

Yik, M., Russell, J. A., & Steiger, J. H. (2011). A 12-point circumplex structure of core affect. *Emotion*, *11*(4), 705–731. https://doi.org/10.1037/a0023980.

Cite this article: Fahrenwaldt, A., tho Pesch, F., Fiedler, S., and Baumert, A. (2024). What's moral wiggle room? A theory specification. *Judgment and Decision Making*, e17. https://doi.org/10.1017/jdm.2024.16