Cognition 239 (2023) 105579

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Why do evaluative judgments affect emotion attributions? The roles of judgments about fittingness and the true self

Michael Prinzing [a,*], Brian Earp [b], Joshua Knobe [a]

[a] Yale University, United States
[b] Oxford University, United States

ARTICLE INFO

Keywords:
True self
Moral cognition
Emotion concepts
True happiness
True love

ABSTRACT

Past research has found that the value of a person's activities can affect observers' judgments about whether that person is experiencing certain emotions (e.g., people consider morally good agents happier than morally bad agents). One proposed explanation for this effect is that emotion attributions are influenced by judgments about *fittingness* (whether the emotion is merited). Another hypothesis is that emotion attributions are influenced by judgments about the agent's *true self* (whether the emotion reflects how the agent feels "deep down"). We tested these hypotheses in six studies. After finding that people think a wide range of emotions can be fitting and reflect a person's true self (Study 1), we tested the predictions of these two hypotheses for attributions of happiness, love, sadness, and hatred. We manipulated the emotions' fittingness (Studies 2a-b and 4) and whether the emotions reflected an agent's true self (Studies 3 and 5), measuring emotion attributions as well as fittingness judgments and true self judgments. The fittingness manipulation only impacted emotion attributions in the cases where it also impacted true self judgments, whereas the true self manipulation impacted emotion attribution in all cases, including those where it did not impact fittingness judgments. These results cast serious doubt on the fittingness hypothesis and offer some support for the true self hypothesis, which could be developed further in future work.

Imagine a nurse who is almost always in a pleasant mood and who feels very satisfied with her life. Now suppose you learn that she has been intentionally poisoning her patients. Is she happy? A large body of research has found that this kind of information about the value of a person's activities influences people's judgments about happiness (Díaz & Reuter, 2020; Kneer & Haybron, 2023; J. Phillips, De Freitas, Mott, Gruber, & Knobe, 2017; Prinzing & Fredrickson, 2022). Even when a person is explicitly described as experiencing lots of pleasant emotion, little unpleasant emotion, and as being very satisfied with life, people are less inclined to think that the person is happy when they judge the person's activities to be immoral, shallow, or worthless.

This sort of effect is strikingly robust and widespread. It has been observed in young children and across cultures and languages (Chen, Harris, & Yang, 2023; Yang, Knobe, & Dunham, 2020) and even among academics specializing in happiness research (J. Phillips et al., 2017). The effect also extends beyond the case of happiness specifically. For instance, prior studies have found similar effects on attributions of love (J. Phillips, Misenheimer, & Knobe, 2011) and attributions of "pleasant emotions" and "unpleasant emotions" in general (Prinzing & Fredrickson, 2022). Naturally, evaluative considerations do not lead people to completely deny that someone is experiencing an emotion when they otherwise appear to be. For example, people don't typically think that the immoral nurse is *un*happy. Instead, evaluative considerations make people hesitant or uncertain in their attributions of emotions. Although there is clearly a sense in which the immoral nurse feels happy, there is also something amiss about her happiness.

In this paper, we test two hypotheses that purport to explain the influence of evaluative considerations on attributions of emotions. These two hypotheses offer very different explanations and are based on separate bodies of research, yet prior studies have offered support for each. We attempt to pit these two hypotheses against each other by examining their implications across a range of different emotions.

1. Two candidate hypotheses

One candidate hypothesis draws on the philosophical literature on emotion concepts. Within that literature, it is commonly argued that emotions can be unfitting—meaning that they are not merited or "called

* Corresponding author.
E-mail address: michael@prinzing.net (M. Prinzing).

https://doi.org/10.1016/j.cognition.2023.105579
Received 14 December 2022; Received in revised form 24 July 2023; Accepted 25 July 2023
Available online 29 July 2023

for" by the circumstances (D'Arms & Jacobson, 2000; Howard, 2018). For example, it would be unfitting to feel fear if the object of one's fear is not actually dangerous, to feel hatred if the person one hates is perfectly kind and wonderful, or to feel happy if one is living a bad life (Nozick, 1989; Tatarkiewicz, 1976). Building on this philosophical theory, Díaz and Reuter (2020) have proposed an empirical claim about ordinary judgments, which we call the *fittingness hypothesis*. This hypothesis states that people will be somewhat reluctant to judge that a person is experiencing a particular emotion when they judge that the emotion is unfitting. Indeed, their studies supported this hypothesis. For example, they found that people were somewhat hesitant to attribute fear to someone who was described as "trembling… distressed and nervous" when that person's situation was not at all dangerous (Díaz & Reuter, 2020).

Within the philosophical literature, some argue that, when a person's emotion is unfitting, although the person is clearly experiencing that emotion, they are not experiencing the emotion in its truest or most complete form (M. W. Martin, 2012; Warburton, 2011). One way of putting this point is to say that a person might experience fear, hatred or happiness even when the emotion is not fitting, but that what the person experiences is not "true fear," "true hatred," or "true happiness" (De Sousa, 2002; Hamlyn, 1989; Salmela, 2006). This idea might explain the effect observed in cases like that of the immoral nurse. The immoral nurse might feel happy, but her happiness is unfitting. So, if participants think that the most complete and profound form of happiness requires fittingness, they should think that the immoral nurse is not experiencing *true* happiness.

A second hypothesis draws on a popular distinction between how a person feels "on the surface" and how they feel "deep down." Employing this distinction, we can offer a fairly straightforward account of why participants might be reluctant to say that a person is experiencing a particular emotion, even when the person is explicitly described as having all the usual thoughts and feelings associated with that emotion. The core idea would be that, even if people are clearly and explicitly told that someone is experiencing the relevant sorts of feelings, they might, nonetheless, believe that the person has very *different* feelings at a deeper level (Morgan & Averill, 1992; Prinzing & Fredrickson, 2022). A growing body of research interprets claims about what a person is like deep down as claims about the person's "true self" (Christy, Schlegel, & Cimpian, 2019; De Freitas, Cikara, Grossmann, & Schlegel, 2017; Hicks, Schlegel, & Newman, 2019; Strohminger, Knobe, & Newman, 2017). Accordingly, we refer to this second hypothesis as the *true self hypothesis*.

The true self hypothesis does not deny that people will sometimes say a person is experiencing an emotion even when the emotion does not reflect the person's true self. Rather, it states that, if people judge that a person's emotion does not reflect the person's true self, then people may think that the person is not experiencing the emotion in the most complete and profound way. Once again, a natural way to express this idea might be with the modifier "true." Defenders of the true self hypothesis might predict that people won't completely deny that a person is experiencing happiness or hatred when these emotions fail to reflect the person's true self, but that people will think that such a person is not experiencing "true happiness" or "true hatred."

This second hypothesis provides a very different explanation for participants' judgments in cases like that of the immoral nurse. When a person does something that is immoral, participants may think that although she feels happy on the surface, deep down she is feeling something very different. Past research has provided some evidence for this claim. One study found that, when people read about a person whose life is morally bad but who nonetheless feels good all the time, they tend to conclude the person feels differently "deep down" versus "on the surface" (Newman, Freitas, & Knobe, 2015). Another study similarly found that people think such the person does not experience a lot of pleasant feelings and satisfaction "deep down" and is not at peace with herself (Prinzing & Fredrickson, 2022).

In short, the fittingness hypothesis and the true self hypothesis provide very different explanations for the effect of evaluative considerations on attributions of emotion. But, in the particular case where a person does something morally bad and participants are asked whether that person is happy, the two hypotheses make exactly the same prediction. The fittingness hypothesis predicts that people will judge that what the person feels is not true happiness because the person's feelings are unfitting. Similarly, the true self hypothesis predicts that people will judge that what the person feels is not true happiness because the person's feelings do not reflect her true self. Since the two hypotheses make exactly the same prediction in cases of this kind, it might be difficult to use such cases to adequately test the two hypotheses.

If we want a better test, it might therefore be helpful to expand the scope of our inquiry and look at a broader range of emotions. For example, we can look at judgments about sadness, love, or hatred. When considering these other emotions, we may find that fittingness judgments and true self judgments do not track each other as closely. Thus, we may find cases in which an experimental manipulation affects judgments about whether an emotion is fitting but *not* judgments about whether the emotion reflects the true self, or vice versa. In such cases, the two hypotheses will make very different predictions.

2. Which emotions can reflect the true self?

This strategy will only be feasible if people think that a range of different emotions could potentially reflect a person's true self. If people only apply the notion of a true self to a relatively small handful of emotions, then the approach we propose to employ here would not make sense. We therefore need to begin by investigating whether people actually think of a broad range of different emotions as potentially reflecting the true self.

On one hand, there is some evidence suggesting that people do not think of the true self in that way. In particular, previous research suggests that people tend to believe that a person's true self calls that person to do what is *morally good*. For instance, adults and children alike consider moral attributes to be essential to the self (Heiphetz, Strohminger, Gelman, & Young, 2018), and people are more inclined to consider morally good traits to be essential than morally neutral or bad traits (Zhang & Alicke, 2021). Other studies have found that people are more morally motivated when they reflect on their true selves (Kim, Christy, Rivera, Schlegel, & Hicks, 2018), and tend to see good behavior as a reflection of a person's true self, whereas they see bad behavior as a deviation from the true self (Newman, Bloom, & Knobe, 2014). This asymmetry in people's beliefs about good and bad behavior has emerged in numerous cultures (De Freitas et al., 2018), and even when people think about members of potentially threatening out-groups (De Freitas & Cikara, 2018). Hence, it may be that people think only deeply valuable emotions, such as happiness or love, can reflect one's true self. Whereas they might think of negative emotions, such as hatred, as arising from something other than the true self.

However, other studies have found that people sometimes think certain kinds of morally bad states and behaviors reflect a person's true self. For example, people sometimes essentialize criminality—i.e., they think of criminal behaviors as revealing something about the essence of the agents who engage in those behaviors (J. W. Martin, Charles, & Heiphetz, 2022). People also judge certain kinds of morally bad behavior, such as blatantly racist behavior, to reflect a person's true self, with consequences for judgments of blameworthiness (Daigle & Demaree-Cotton, 2022). Finally, when participants are told that an agent has a morally bad true self, participants seem willing to accept this description, with downstream effects on numerous judgments (Earp, Skorburg, Everett, & Savulescu, 2019; Newman et al., 2015).

Moreover, other research has identified a range of factors, totally unrelated to morality, that can influence people's judgments about the true self. People are more inclined to see a psychological state as reflecting the true self when it is stable and immutable (Christy et al., 2019), when it is found even at the end of the agent's life (Earp,

Hannikainen, Dale, & Latham, 2023; Newman, Lockhart, & Keil, 2010), and when it is grounded in intuition rather than careful reflection (Maglio & Reich, 2019; Oktar & Lombrozo, 2022). Thus, even if there is a very strong main effect such that people are more inclined to regard a mental state as reflecting the agent's true self when it is morally good than when it is morally bad, it might still be possible for morally bad states and states unrelated to morality to be seen as reflecting the true self.

In short, it is an open empirical question whether people see emotions like sadness and hatred as the sorts of things that can reflect a person's true self. If that is not the case, then the strategy that we have proposed for pitting the fittingness hypothesis and true self hypothesis against each other is not feasible. However, it is also possible that people think that emotions like sadness and hatred can reflect the true self, and if they do, examining judgments across a range of very different emotions might offer us some helpful insight regarding the two hypotheses.

3. The present studies

To test the fittingness and true self hypotheses, we conducted a series of studies that examined people's judgments across several different emotions. In each study, we used a manipulation designed to influence people's judgments about whether the emotion was fitting or unfitting or whether it did or did not reflect the agent's true self. Then we asked whether the impact of the manipulations on people's emotion attributions mirrored the impact on their fittingness judgments or whether it mirrored the impact on their true self judgments.

As an illustration of the basic logic of this approach, consider a study in which participants are given vignettes about agents experiencing various emotions and in which we manipulate the fittingness of each emotion. That is, participants read about fitting happiness or unfitting happiness, fitting sadness or unfitting sadness, and so forth. Suppose that, in some cases, when participants read about an unfitting emotion, they tend to conclude that the emotion does not reflect the agent's true self. In other cases, this effect on true self judgments does not emerge. If this were the result, then the two hypotheses would generate very different predictions about how people will attribute these emotions. The fittingness hypothesis would predict that the manipulation would impact emotion attributions in a way that follows the impact on fittingness judgments—i.e., there should be a straightforward main effect whereby emotion attributions are always lower when the emotion is seen as less fitting. By contrast, the true self hypothesis would predict that the manipulation will impact emotion attributions in a way that follows the impact on true self judgments—i.e., there should be a complex pattern whereby the emotion attributions are only lower in cases where true self judgments are also lower. These effects on emotion attributions need not be exactly the same size as the effects on the fittingness or true self judgments. But, for the predictions of these hypotheses to be supported, emotion attributions need to be impacted in the same direction as the corresponding judgment.

We pursued this strategy in six studies. In Study 1, we examined a wide range of different emotions, testing whether people think that it makes sense to say that the emotion is fitting, reflects a person's true self, and is "true." In each of the studies that followed, we focused on four specific emotions (happiness, love, sadness, and hatred), presenting participants with vignettes about agents who appeared to be experiencing a particular emotion and examining participants' judgments. In Studies 2a-b and 4 we manipulated whether the emotions were fitting or unfitting. In Studies 3 and 5, we manipulated whether the agents' emotions reflected their true selves (i.e., we described the agent's feelings as either arising from deep down inside or as being only on the surface).

Whereas, in Studies 2a-b and 3, we assessed emotion attributions in the typical way, in Studies 4 and 5, we assessed participants' willingness to attribute "true" happiness, love, sadness, and hatred. The idea was that, if the effects observed on regular emotion attributions arise because people think that the agents are only experiencing these emotions in a limited and mundane sense, and not in a more complete and profound sense, then the effects of our experimental manipulations should be magnified when the measures specifically ask whether the emotions are "true."

Prior to conducting these experiments, we did not have any expectation about which of the two candidate hypotheses, if either, was correct. In fact, different authors of the present work have previously argued for different hypotheses (Newman et al., 2015; J. Phillips et al., 2017; Prinzing & Fredrickson, 2022). The aim was to give each hypothesis the best chance at explaining the data, recognizing that each might be, at least partially, correct. All data, materials, analytic code, and pre-registration forms are available online: https://osf.io/fvbh6/.

4. Study 1

Based on the existing literature, we have speculated that people think a wide range of different emotions can be fitting or unfitting, can reflect a person's true self, and can be called "true." This study sought to test these speculations by investigating people's intuitive judgments about 38 different emotions. Of particular interest is the question of whether people think that negative emotions can reflect a person's true self. On the one hand, past research has found that people generally think of the true self as morally good. This could indicate that only certain positive emotions will be thought to reflect one's true self. However, other research suggests that people's ordinary true self beliefs are very complex and that, at least in some cases, people do think that negative emotions can reflect a person's true self. This study enables us to address this question.

4.1. Method

Participants. We recruited two samples of participants using Prolific, an online research platform. We recruited Sample 1 first, pre-registering a plan to examine ratings of goodness and trueness for 38 different emotions. We then separately pre-registered a plan to recruit Sample 2 to examine ratings of fittingness and true self for the same emotions. In the present study, we combine the data from the two samples.

For Sample 1, we received 402 complete responses. We included an attention check in the survey—a question that read: "For this question please simply place the slider at 65." As pre-registered, we excluded participants ($n = 7$) who did not mark 65, leaving $N = 395$ ($M_{age} = 40.28$, $SD_{age} = 16.14$; 43.3% men, 56.2% women, < 1% other gender or prefer not to say; 6.3% Asian or Asian American, 25.3% Black or African American; 4.1% Hispanic or Latinx, 55.4% White or European American, 4.6% other race, 4.1% mixed race, < 1% prefer not to say).

For Sample 2, we received 401 complete responses. As a quality-control check, at the end of the survey, we asked participants, "Which day of the week comes after Tuesday and before Thursday?" We pre-registered that we would exclude any participants who did not enter "Wednesday" (minor typos permitted). However, none of the participants failed this check. We therefore include all $N = 401$ participants ($M_{age} = 37.04$, $SD_{age} = 12.90$; 48.1% identified as men, 49.4% women, and 2.5% other gender or prefer not to say; 11.0% identified as Asian, 7.5% Black or African American, 6.2% Hispanic or Latinx/Latiné, 66.6% White, 7.2% mixed race, 1.5% other race or prefer not to say).

Procedure and Measures. All participants were randomly assigned to answer a single question for 20 emotions, randomly drawn from a pool of 38 emotions. The response scales for all questions were 100-point slider scales (anchors given in parentheses below). For Sample 1, participants were randomly assigned to answer one of the following four questions:

- Goodness 1: "All else being equal, to the extent that a person feels __, this will make their life go…" (0 = "Much worse", 100 = "Much better")

- Goodness 2: "When thinking about __, would you say that feeling this emotion is generally a negative or positive thing?" (0 = "Extremely negative", 100 = "Extremely positive")
- Trueness 1: "Please indicate the extent to which each of the following statements sound weird or natural. Could you imagine saying or hearing someone say each of these statements in a normal conversation? That person is experiencing *true* __." (0 = "Sounds weird", 100 = "Sounds natural")
- Trueness 2: "Please indicate the extent to which each of the following statements makes sense. If someone said each of these statements in a real-life conversation, would the statements even make sense? Imagine someone says: 'I suppose there is a sense in which that person is feeling __. But it isn't *true* __.' Does this statement even make sense?" (0 = "Makes no sense at all", 100 = "Makes perfect sense")

Sample 2 participants were randomly assigned to answer one of the following two questions:

- Fittingness: "When people talk about emotions, sometimes they will say that a certain emotion is fitting or merited—i.e., when you recognize the way things really are, it just makes sense to feel that emotion. In some cases, these kinds of claims seem perfectly natural. In other cases, they might seem weird. For each of the following emotions, please indicate how weird or natural the statement sounds. This person's __ makes sense and is merited." (0 = "Sounds weird" to 100 = "Sounds natural")
- True Self: "When people talk about emotions, sometimes they will say that an emotion reflects a person's true self—i.e., it expresses who a person really is. In some cases, these kinds of claims seem perfectly natural. In other cases, they might seem weird. For each of the

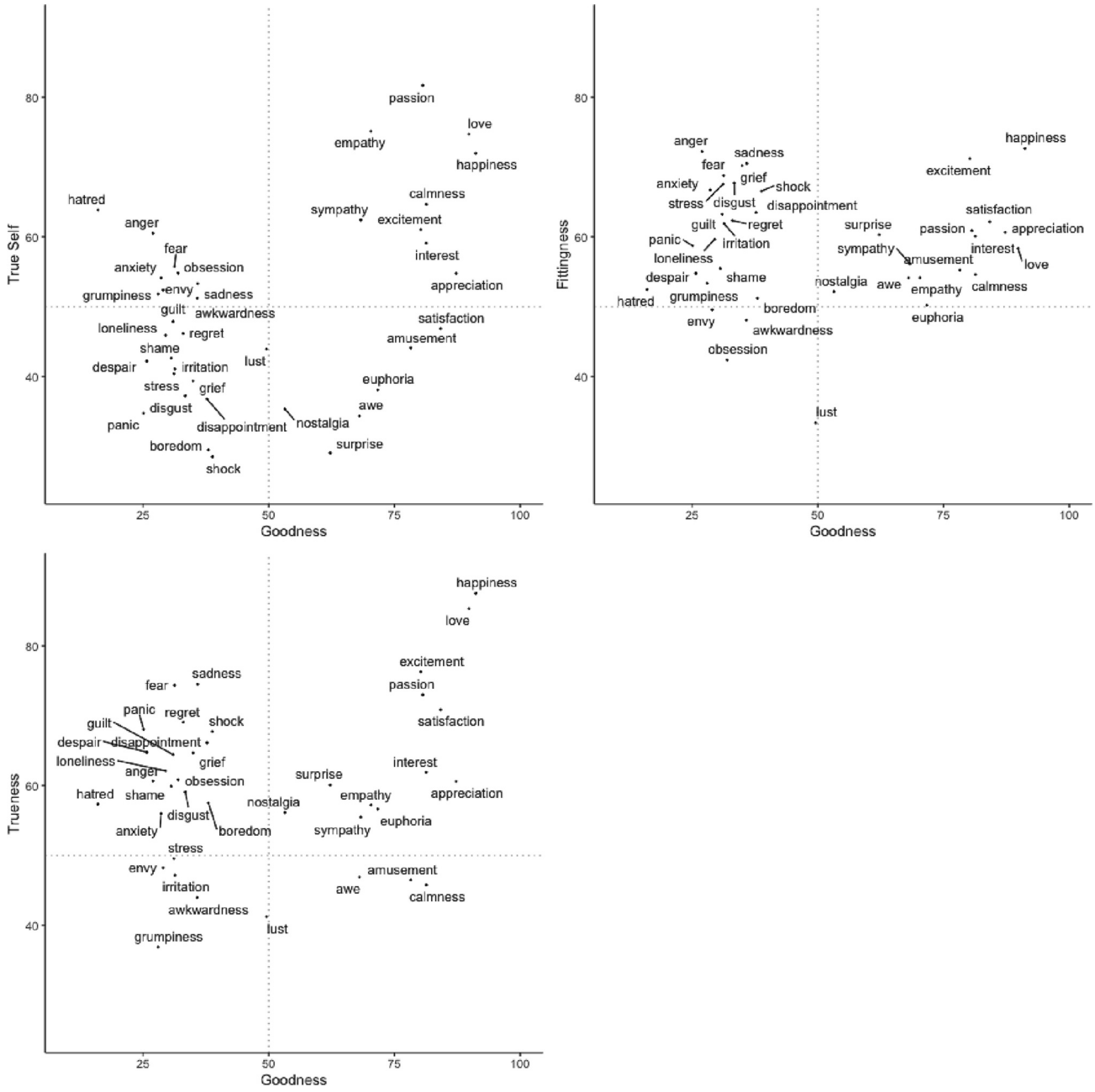

Fig. 1. Scatterplots of emotions, plotted according to the means for goodness and each of the other three question types. Dotted grey lines indicate the midpoints on each scale.

following emotions, please indicate how weird or natural the statement seems. This person's __ is a reflection of her true self, the person she really is."

4.2. Results and discussion

We computed the means of responses to each question for each emotion. As pre-registered, we then averaged the two goodness questions ($r = 0.99$, $p < .001$), and the two trueness questions ($r = 0.61$, $p < .001$). Fig. 1 shows scatterplots of these means, plotting the means for trueness ($M_{grand} = 60.38$, $SD = 11.49$), fittingness ($M_{grand} = 59.04$, $SD = 8.59$) and true self ($M_{grand} = 49.67$, $SD = 13.49$) against the means for goodness ($M_{grand} = 49.50$, $SD = 23.59$).

The results for fittingness indicated that, with a few exceptions (such as lust and obsession) people generally think that most emotions can be fitting. For example, people think that a person's happiness, sadness, excitement, and anger can each make sense and be merited. The results for trueness indicated that, although people think it does not make sense to apply the word "true" to certain emotions (e.g., "true lust" or "true grumpiness"), it does make sense to apply "true" to a range of different emotions. Moreover, the emotions that people think can be called "true" include both positive and negative emotions. People think it makes perfect sense to talk of "true happiness" and "true love" (Earp, Do, & Knobe, 2021; Lantian, Boudesseul, & Cova, 2023), for instance, but also "true sadness" (Lopez, 1974) and "true fear".

Importantly, the results on the true self question indicate that people think a range of different emotions can reflect the true self. Whereas people tended not to think that shock or boredom could reflect the true self, they did think that emotions like happiness and love, but also hatred and sadness, can each reflect the true self. There appears to be a complex relationship between the degree to which people think an emotion is good and the degree to which they think that it can reflect the true self. The emotions with the highest means on the true self question were very positive emotions (passion, empathy, and love). However, the emotions with the lowest means were not very negative emotions, but neutral ones like shock and surprise. In fact, a number of very negative emotions, including hatred, fear, and anger, had relatively high means. Hence, although people think that the emotions that most clearly reflect the true self are certain positive emotions, they don't think that *only* positive emotions can reflect one's true self.

Future work could investigate why some emotions are thought to reflect a person's true self more than others. One hypothesis would be that people think of an emotion as reflecting a person's true self to the degree that they think it reveals the person's values. Past work has found that, when someone's values change substantially, people are inclined to say that the person is no longer the same person that they were before (Hitlin, 2003; Strohminger & Nichols, 2014). Given that a person's values are central to ordinary thinking about personal identity, future work might fruitfully investigate whether a person's values are also central to ordinary thinking about the true self.

In summary, the results of this first study suggest that people think a variety of different emotions, both positive and negative, can be fitting, reflect one's true self, and be called "true." The question then becomes whether we can identify cases in which people's judgments about fittingness and the true self diverge. By examining such cases, we can determine which factor influences attributions of emotions.

5. Study 2a

In this study, we presented participants with vignettes about agents who appeared to be experiencing happiness, love, sadness, and hatred. We manipulated whether these emotions would be fitting or unfitting and examined the effects on judgments about whether the agent was experiencing the relevant emotion. We also assessed participants' own judgments about whether the emotion is fitting and whether it reflects the agent's true self. If the manipulation works as intended (i.e., if it reduces fittingness judgments for all four emotions) then the fittingness hypothesis predicts reduced attributions for all four emotions. The true self hypothesis, by contrast, does not predict reduced emotion attributions except in cases where the unfittingness manipulation incidentally reduces true self judgments.

Because we assessed both fittingness and true self judgments, we were also able to investigate whether each kind of judgment predicts emotion attributions after controlling for experimental condition. That is, we can further test these two hypotheses by comparing the proportion of the variance in emotion attributions that is explained by fittingness judgments versus true self judgments.

5.1. Method

Participants. We recruited 608 participants from Prolific, an online research platform. As pre-registered, we excluded from analyses participants ($n = 8$) who did not correctly answer a quality-check question at the end of the survey ("What day of the week is it?"). This left $N = 600$ participants ($M_{age} = 35.16$, $SD_{age} = 11.92$; 48.5% identified as men, 49.5% women, and 2.0% other gender or prefer not to say; 8.8% identified as Asian, 4.7% Black or African American, 5.7% Hispanic or Latinx/Latiné, 71.2% White, 8.8% mixed race, < 1% other race or prefer not to say).

Procedure. Participants read short vignettes featuring an agent named Mario. We randomized participants to one of eight vignettes in a 2 (unfittingness manipulation: unfitting, fitting) x 4 (emotion: happiness, sadness, love, and hatred) experimental design. In the happiness, sadness, love, and hatred conditions, Mario was described, respectively, as: satisfied with his life and in a cheerful mood most of the time; dissatisfied with his life and in a gloomy mood most of the time; having very romantic feelings and a sense of intimacy with a woman named Dax; and as having very antagonistic and hostile feelings towards a woman named Dax.

We manipulated the fittingness of each emotion by altering the descriptions of the circumstances of Mario's life. For happiness, Mario was described either as "the father of three children who all really love him" (fitting) or as being "so preoccupied with becoming popular that he no longer bothers to stay in touch with his old friends unless they know someone famous" (unfitting). For sadness, Mario was described as an artist who either "has never sold a single painting" and in whom "potential buyers never seem to take an interest" (fitting) or as having "sold hundreds of paintings" and in whom "potential buyers always seem to take an interest" (unfitting). For love, Dax was described as having "an insightful, imaginative intellect," a "great sense of humor," and as being "totally honest and reliable" (fitting), or as having "a 'walking encyclopedia' kind of intellect," "no sense of humor at all," and as being "totally dishonest and unreliable" (unfitting). In the hatred conditions, Dax was described as a political activist who either has "connections with a couple of White supremacist groups" and works "to undermine voting rights for minorities" (fitting), or who has "connections with a couple of civil rights groups" and works "to defend voting rights for minorities" (unfitting).

The descriptions of the agent's life circumstances always came before the descriptions of the agent's feelings. Although past research indicates that presentation order should not influence the results (J. Phillips et al., 2017), this reflects a potential limitation of these materials.

Measures. We assessed emotion attributions by asking participants to indicate the extent to which they agreed with the statement, "Mario __", where the blank was filled with "is happy", "is sad", "loves Dax," or "hates Dax," according to condition. The response scale for this and the following items was a 100-point slider, anchored with "Not at all" and "Absolutely." On the following page, we assessed fittingness and true self judgments as follows:

"Now we want to ask you two different questions about the story you read.

Sometimes [emotion] is merited. When you think about [someone /

your life], it totally makes sense to feel [emotion]. Other times, it isn't merited. When you recognize what [someone / your life] is actually like, it just doesn't make any sense why you would feel [emotion]. This question is about [Dax / Mario's life]. Do you think that [Dax / Mario's life] merits Mario's [emotion]?

Sometimes people feel [emotion] deep down. The feeling seems to express who a person, in some sense, really is. Other times, [emotion] is more "surface level" and doesn't say much about a person's true self. This question is about Mario's feelings. Do you think Mario's [emotion] is a reflection of his true self?"

Brackets indicate text that varied across conditions. In the happiness and sadness conditions, the fittingness question asked about "your life" and "Mario's life." In the love and hatred conditions, it asked about "someone" and "Dax." We randomized the order in which these two questions were presented.

5.2. Results

Fig. 2 shows the means and distributions for fittingness judgments, true self judgments, and emotion attributions across conditions. The key question in this study was whether the effects on emotion attributions (the third column in Fig. 2) would look more like the effects on fittingness judgments (first column), as the fittingness hypothesis predicts, or if they would look more like the effects on true self judgments (second column), as the true self hypothesis predicts. To examine these effects, we ran three factorial ANOVAs, one for each dependent variable.

For fittingness judgments, there was a very large main effect of the manipulation, $F(1, 592) = 476.49$, $p < .001$, $\eta_p^2 = 0.45$ [0.39, 0.50], as well as a main effect of emotion, $F(3, 592) = 16.66$, $p < .001$, $\eta_p^2 = 0.08$ [0.04, 0.12], and an interaction, $F(3, 592) = 16.92$, $p < .001$, $\eta_p^2 = 0.08$ [0.04, 0.12]. We decomposed the interaction by examining simple main effects, using the "emmeans" function from the *emmeans* package (Lenth, Singmann, Love, Buerkner, & Herve, 2018). This revealed that, although the size of the manipulation's effect varied somewhat across emotions, it was significant for each (all $ps < 0.001$). Hence, the manipulation performed as expected, substantially reducing participants' judgments about the fittingness of each emotion.

For true self judgments, there was no main effect of the manipulation, $F(1, 592) = 1.26$, $p = .261$, $\eta_p^2 = 0.00$ [0.00, 0.02], but there was a main effect of emotion, $F(3, 592) = 6.69$, $p < .001$, $\eta_p^2 = 0.03$ [0.01, 0.06], and an interaction, $F(3, 592) = 38.06$, $p < .001$, $\eta_p^2 = 0.16$ [0.11, 0.21]. Decomposing the interaction revealed that the manipulation reduced true self judgments for happiness, $t(592) = -7.82$, $p < .001$, and love $t(592) = -3.09$, $p = .002$, but it *increased* true self judgments for

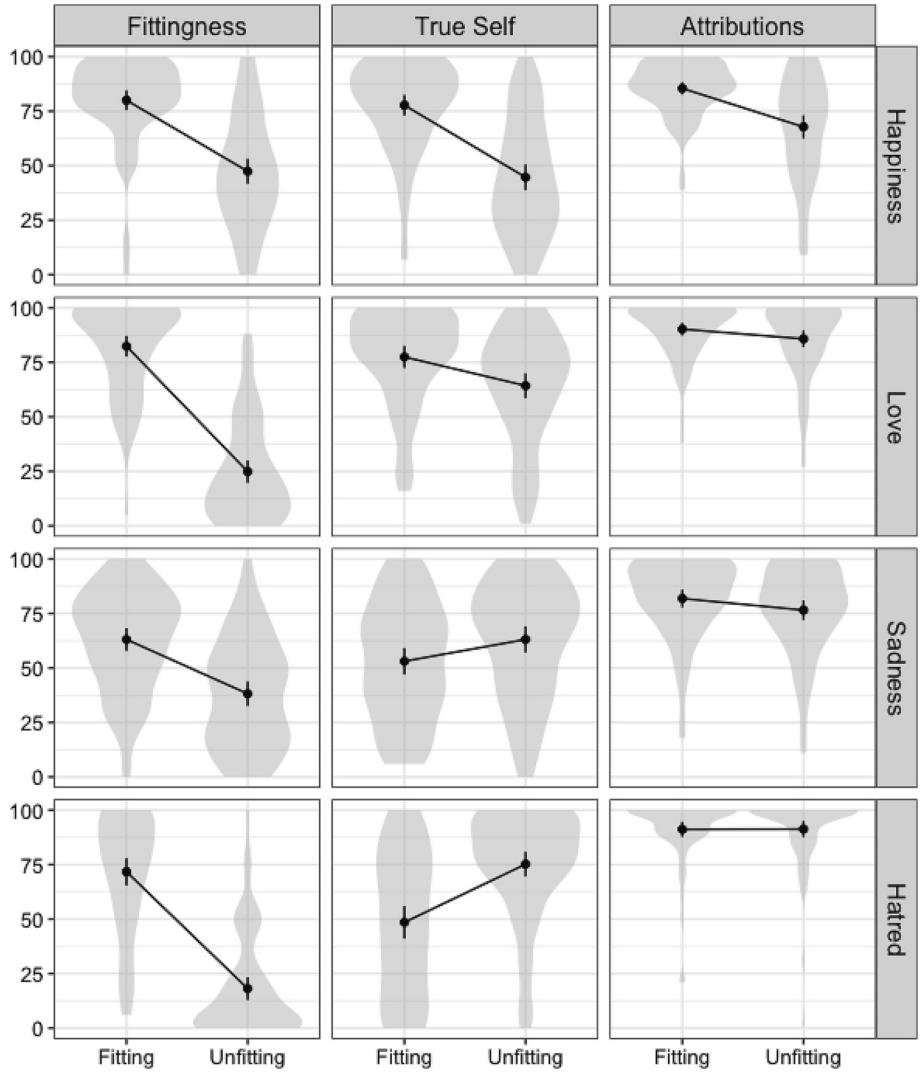

Fig. 2. Violin Plots for Dependent Variables Across Experimental Conditions in Study 2a | Points indicate condition means. Error bars indicate 95% confidence intervals.

Table 1
ANCOVA Results for all studies | Bracketed ranges indicate 95% confidence intervals. * indicates $p < .05$; ** indicates $p < .01$; *** indicates $p < .001$.

	DV: Emotion Attributions								**DV: True Emotion Attributions**			
	Study 2a				**Study 2b**				**Study 4**			
	Model 1		Model 2		Model 1		Model 2		Model 1		Model 2	
Predictor	*F*	η_p^2	*F*	η_p^2	*F*	η_p^2	*F*	η_p^2	*F*	η_p^2	*F*	η_p^2
Emotion	22.51***	0.10 [0.06, 0.15]	21.86***	0.10 [0.06, 0.15]	7.22***	0.04 [0.01, 0.07]	8.76***	0.04 [0.01, 0.08]	22.28***	0.10 [0.06, 0.15]	25.88***	0.12 [0.07, 0.16]
Unfittingness Manipulation	21.91***	0.04 [0.01, 0.07]	7.20**	0.01 [0.00, 0.04]	32.75***	0.05 [0.02, 0.09]	7.55**	0.01 [0.00, 0.04]	72.21***	0.11 [0.07, 0.16]	11.85***	0.02 [0.00, 0.05]
Emotion x Manipulation	2.20	0.01 [0.00, 0.03]	2.59	0.01 [0.00, 0.03]	1.70	0.00 [0.00, 0.03]	1.98	0.00 [0.00, 0.03]	9.26***	0.04 [0.02, 0.08]	11.24***	0.05 [0.02, 0.09]
True Self Judgments	43.59***	0.07 [0.03, 0.11]	38.81***	0.06 [0.03, 0.10]	49.90***	0.08 [0.04, 0.12]	49.19***	0.08 [0.04, 0.12]	166.39***	0.22 [0.17, 0.28]	145.03***	0.20 [0.14, 0.25]
Fittingness Judgments	–	–	1.38	0.00 [0.00, 0.02]	–	–	8.60**	0.01 [0.00, 0.04]	–	–	17.04***	0.03 [0.01, 0.06]
	Study 3								**Study 5**			
	Model 1		Model 2						Model 1		Model 2	
Predictor	*F*	η_p^2	*F*	η_p^2					*F*	η_p^2	*F*	η_p^2
Emotion	7.41***	0.04 [0.01, 0.07]	15.81***	0.07 [0.04, 0.12]					2.57	0.01 [0.00, 0.03]	9.77***	0.05 [0.02, 0.08]
True Self Manipulation	499.48***	0.46 [0.41, 0.51]	243.96***	0.29 [0.24, 0.35]					455.99***	0.44 [0.38, 0.49]	250.98***	0.30 [0.24, 0.35]
Emotion x Manipulation	16.25***	0.08 [0.04, 0.12]	12.07***	0.06 [0.02, 0.10]					6.19***	0.03 [0.01, 0.06]	7.02***	0.03 [0.01, 0.06]
True Self Judgments	–	–	67.54***	0.10 [0.06, 0.15]					–	–	112.81***	0.16 [0.11, 0.21]
Fittingness Judgments	31.97***	0.05 [0.02, 0.09]	13.47***	0.02 [0.00, 0.05]					86.08***	0.13 [0.08, 0.18]	43.18***	0.07 [0.03, 0.11]

sadness, $t(592) = 6.27$, $p < .001$, and hatred, $t(592) = 2.37$, $p = .018$.

Thus, the experimental manipulation reduced fittingness judgments for all four emotions, but only reduced true self judgments for happiness and love. This difference enabled us to test whether the effects on emotion attributions look more like the effects on fittingness judgments (i.e., reductions across the board) or more like the effects on true self judgments (i.e., reductions for happiness and love and increases for sadness and hatred).

For emotion attributions, there was a main effect of the unfittingness manipulation, $F(1, 592) = 23.24$, $p < .001$, $\eta_p^2 = 0.04$ [0.01, 0.07], as well as emotion, $F(3, 592) = 24.15$, $p < .001$, $\eta_p^2 = 0.11$ [0.06, 0.16], and an interaction, $F(3, 592) = 7.21$, $p < .001$, $\eta_p^2 = 0.04$ [0.01, 0.07]. Decomposing the interaction revealed that the manipulation reduced attributions of happiness, $t(592) = -6.25$, $p < .001$, $d = -1.02$, 95% CI: [−1.34, −0.69]. However, it did not significantly affect attributions of sadness, $t(592) = -1.89$, $d = -0.31$, $p = .060$, 95% CI: [−0.63, 0.01], love, $t(592) = -1.57$, $p = .117$, $d = -0.26$, 95% CI: [−0.58, 0.06], or hatred, $t(592) = 0.04$, $p = .967$, $d = 0.00$, 95% CI: [−0.32, 0.33].

Finally, instead of comparing means across conditions, we shifted to examining the data at the participant level. We used ANCOVAs to test for unique associations between these two kinds of judgments and emotion attributions, after controlling for experimental condition (see Table 1 for results). The aim was to determine whether each kind of judgment independently predicted emotion attributions and, if so, how much variance in those attributions each explained. In Model 1, alongside the experimental factors and their interaction, we included a term for participants' true self judgments. This term proved significant. In fact, it accounted for at least as much variance in emotion attributions ($\eta_p^2 = 0.07$) as the unfittingness manipulation itself ($\eta_p^2 = 0.04$). In Model 2, we also added a term for participants' fittingness judgments. This term did not reach statistical significance, though it was close ($p = .052$). Moreover, adding it to the model did not substantially reduce the variance explained by true self judgments ($\Delta\eta_p^2 = 0.01$). In other words, true self judgments appear to be uniquely and strongly associated with emotion attributions, whereas fittingness judgments are not.

5.3. Discussion

Our strategy for testing the fittingness and true self hypotheses involves looking across a range of different emotions to find cases in which people's judgments about whether an agent's emotion is fitting diverge from people's judgments about whether the emotion reflects the agent's true self. In such cases, the two hypotheses make opposite predictions about people's emotion attributions.

As intended, the unfittingness manipulation reduced fittingness judgments for all emotions. Hence, the fittingness hypothesis predicts reduced attributions for all emotions. Yet this is not at all what we observed. Only happiness attributions were significantly reduced. Hence the fittingness hypothesis makes the right prediction only in that one case, and makes the wrong predictions for sadness, love and hatred. Additionally, when we examined the associations between participants' judgments, we found that fittingness judgments were not significantly associated with emotion attributions after controlling for condition and true self judgments. These results speak strongly against the fittingness hypothesis.

Intriguingly, although the unfittingness manipulation said nothing about the agents' true selves, it still influenced true self judgments (see the General Discussion for further discussion of why this might be). The unfittingness manipulation reduced true self judgments for happiness and love but not for sadness or hatred. Hence, the true self hypothesis predicts no reductions in sadness or hatred attributions, and indeed we observed none, whereas the hypothesis does predict reductions for happiness, which we observed, and love, which we did not (though, see Studies 2b and 4). In short, the results suggest that the true self hypothesis has at least some potential to explain the pattern of effects across the different emotions.

Additionally, in examining the associations between participants' judgments, we found that true self judgments significantly predicted emotion attributions after controlling for condition and fittingness judgments. Including true self judgments in this model also substantially reduced the variance in emotion attributions that was explained by the experimental manipulation. This suggests that much of the manipulation's effects on emotion attributions came from its effects on true self judgments. Hence, these results provide some evidence in favor of the true self hypothesis.

However, we also observed an unexpected effect that might seem to speak against the true self hypothesis. Looking at true self judgments for sadness and hatred, we do not simply find that the unfittingness manipulation has no impact. Instead, we find a significant effect in the opposite direction. When participants learned that the agents' sadness and hatred were unfitting, they concluded that these emotions were *more* reflective of the agents' true selves. For example, people thought that an agent's hatred reflected his true self more when he hated a civil rights worker versus a White supremacist. Because the unfittingness manipulation increased true self judgments for sadness and hatred, the hypothesis predicts higher attributions of sadness and hatred in the unfitting conditions, which we did not observe.

6. Study 2b

Study 2a faces an important limitation in that the vignettes included arbitrary differences in the descriptions of the agents' life circumstances. For instance, in the sadness conditions, the agent was described as an artist, whereas in the love conditions he was described as working at a recreational equipment store. It's possible that such differences played some role in shaping the different patterns of results that we observed across emotions. Yet it is difficult to eliminate such differences entirely, as the kinds of life circumstances that make an emotion fitting or unfitting vary across emotions. In this study, we sought to avoid this problem by making the vignettes more abstract. Participants were not given any concrete information about where the agents live, what they do for work, what they spend their time doing, and so on. Instead, participants read highly abstract descriptions. For instance, in the fitting sadness condition, participants read simply that the agent's life was tragic and unfortunate. By not including any concrete details, we ensure that the descriptions of the agents' life circumstances include no arbitrary or idiosyncratic differences across emotion conditions.

Although this approach overcomes a limitation of Study 2a, it has its own downsides. For instance, participants may have more difficulty forming intuitive judgments about abstract scenarios versus concrete ones. Additionally, since the vignettes don't provide substantive details about *why* the agents' emotions are fitting or unfitting, the manipulation might have weaker effects on people's judgments. Hence, if the two studies yielded different patterns of results, the implications would be unclear. However, if we observe the same basic pattern of results in this study as we did in Study 2a, this would be clear evidence that the results reflect generalizable features of the four emotions and not idiosyncrasies of the vignettes from Study 2a.

6.1. Method

Participants. We recruited 603 participants from Prolific, an online research platform. As pre-registered, we excluded from analyses participants ($n = 6$) who did not correctly answer a quality-check question at the end of the survey ("What day of the week is it?"). This left $N = 597$ participants ($M_{age} = 35.68$, $SD_{age} = 12.22$; 49.2% identified as men, 47.2% women, and 3.5% other gender or prefer not to say; 9.2% identified as Asian, 5.4% Black or African American, 6.7% Hispanic or Latinx/Latiné, 70.5% White, 6.9% mixed race, 1.3% other race or prefer not to say).

Procedure. Participants read short descriptions of an unnamed agent. We randomized participants to one of eight vignettes in a 2 (unfittingness manipulation: unfitting, fitting) x 4 (emotion: happiness, love, sadness, hatred) experimental design. The information about the agents' psychological states was the same as in Study 2a. However, in this study, the information relevant to the fittingness of the emotions was entirely abstract. For example, in the happiness conditions, the vignette read: "There are many things that can make someone's life meaningful. Imagine a person whose life embodies [practically all / none] of these things. Her life is incredibly [rich and worthwhile / empty and pointless]." Brackets indicate text that varied depending on whether participants were in the fitting condition (before the slash) or unfitting condition (after the slash). We used similar descriptions for the other emotions. In the sadness conditions, we described the agent's life as being tragic, "cursed and unfortunate" (fitting) or as not being tragic but rather "blessed and fortunate" (unfitting). In the love conditions, we described the agent's significant other as being a good partner, "thoughtful and selfless" (fitting) or as not being a good partner but rather "shallow and selfish" (unfitting). In the hatred conditions, we described the agent's enemy as being immoral, "cruel and vile" (fitting) or as not being immoral but rather "kind and decent" (unfitting).

Measures. We used the same measures as in Study 2a, replacing references to Mario and Dax with "this person," her "significant other," or her "enemy," depending on emotion condition. As before, participants first responded to the emotion attribution question, and then to the fittingness and true self questions (presented in a randomized order).

6.2. Results

Fig. 3 shows the means and distributions for all three variables across conditions. As before, the key question was whether the effects on emotion attributions (the third column in Fig. 3) would look more like the effects on fittingness judgments (first column) or the effects on true self judgments (second column). We ran three factorial ANOVAs to examine the effects of experimental condition on each dependent variable.

For fittingness judgments, there was a very large main effect of the manipulation, $F(1, 589) = 365.47$, $p < .001$, $\eta_p^2 = 0.38$ [0.33, 0.44], and

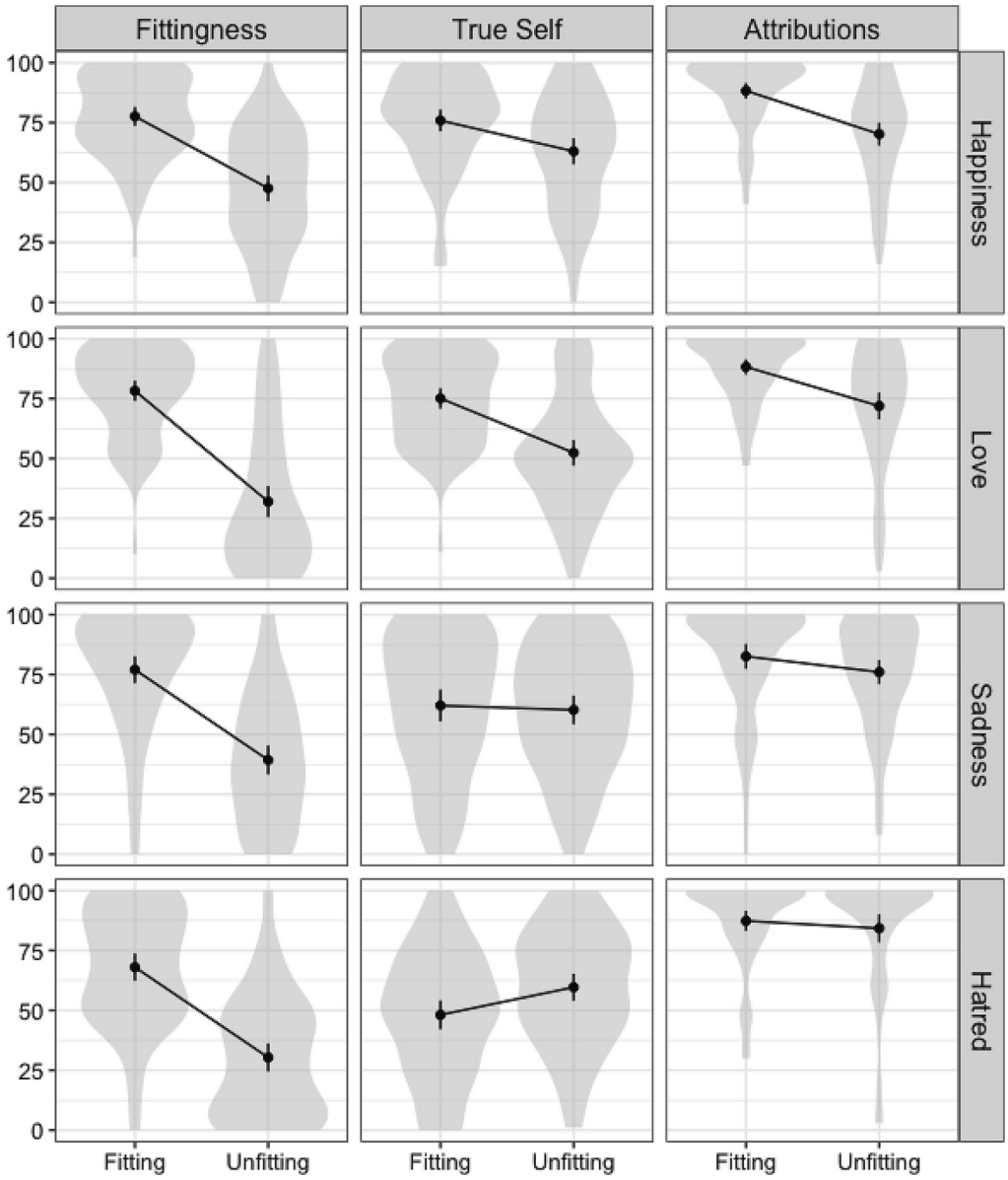

Fig. 3. Violin Plots for Dependent Variables Across Experimental Conditions in Study 2b | Points indicate condition means. Error bars indicate 95% confidence intervals.

emotion, $F(3, 589) = 8.01$, $p < .001$, $\eta_p^2 = 0.04$ [0.01, 0.07], as well as an interaction, $F(3, 589) = 2.80$, $p = .039$, $\eta_p^2 = 0.01$ [0.00, 0.03]. Decomposing this interaction revealed that, although the size of the effect varied somewhat across emotions, it was significant for each (all $ps < 0.001$). Hence, the manipulation performed as expected, substantially reducing participants' judgments about the fittingness of each emotion.

For true self judgments, there were main effects of the manipulation, $F(1, 589) = 10.40$, $p = .001$, $\eta_p^2 = 0.00$ [0.00, 0.02], and emotion, $F(3, 589) = 10.29$, $p < .001$, $\eta_p^2 = 0.05$ [0.02, 0.08], plus an interaction, $F(3, 589) = 13.42$, $p < .001$, $\eta_p^2 = 0.06$ [0.03, 0.10]. Decomposing the interaction revealed that the manipulation reduced true self judgments for happiness, $t(589) = -3.21$, $p = .001$, and love $t(589) = -5.68$, $p < .001$, but it had no effect on true self judgments for sadness, $t(589) = 0.46$, $p = .646$, and increased true self judgments for hatred, $t(589) = 2.86$, $p = .004$. For happiness, love, and hatred, these results are identical to the results of Study 2a. For sadness, however, they are slightly different. Whereas the manipulation previously increased true self judgments for sadness, we now find no effect.

For emotion attributions, there was a main effect of the unfittingness manipulation, $F(1, 589) = 41.59$, $p < .001$, $\eta_p^2 = 0.07$ [0.03, 0.11], as well as emotion, $F(3, 589) = 3.37$, $p = .018$, $\eta_p^2 = 0.02$ [0.00, 0.04], and an interaction effect, $F(3, 589) = 4.58$, $p = .004$, $\eta_p^2 = 0.02$ [0.00, 0.05]. Decomposing the interaction revealed that the manipulation reduced attributions of happiness, $t(589) = -5.28$, $p < .001$, $d = 0.87$, 95% CI: [0.54, 0.1.19], and love, $t(589) = 4.80$, $p < .001$, $d = 0.78$, 95% CI: [0.46, 1.11]. However, it did not significantly affect attributions of sadness, $t(589) = -1.93$, $p = .055$, $d = 0.31$, 95% CI: [−0.01, 0.63], or hatred, $t(589) = 0.90$, $p = .368$, $d = 0.15$, 95% CI: [−0.18, 0.47]. In other words, as in Study 2a, the effects of the manipulation on emotion attributions don't look at all like the effects on fittingness judgments, but they do look somewhat like the effects on true self judgments.

Shifting to the participant-level analyses, we used ANCOVAs to test for unique associations between these two kinds of judgments and emotion attributions. Full results are shown in Table 1. The overall pattern of results was identical to the results from Study 2a, with one exception. Fittingness judgments emerged as significant in Model 2, whereas they did not in Study 2a. Although they were statistically significant, fittingness judgments explained only a minute proportion of the variance in emotion attributions ($\eta_p^2 = 0.01$). Moreover, adding this term to the model did not reduce the variance explained by true self judgments ($\eta_p^2 = 0.08$ in Model 1 and Model 2).

6.3. Discussion

This study sought to replicate the findings of Study 2a using new materials that did not include any arbitrary or idiosyncratic differences in the descriptions of the agents' life circumstances. The idea was that, if the same basic pattern of results emerged again, then this would indicate that the results reflect generalizable features of the way that people think about these four emotions, substantially increasing our confidence in the findings. Indeed, the results of the two studies were extremely similar. In both cases, across four emotions, the unfittingness manipulation had quite different effects on fittingness judgments versus true self judgments. The effects on emotion attributions were more similar to the effects on true self judgments than on fittingness judgments. Furthermore, after controlling for experimental condition, participants' true self judgments explained far more of the variance in their emotion attributions than did their fittingness judgments.

There were three differences between the results of Studies 2a-b. Two of these differences made the results even more favorable for the true self hypothesis, while one made the results slightly more favorable for the fittingness hypothesis. First, we previously observed that the unfittingness manipulation increased true self judgments for sadness but did not affect attributions of sadness. In this study, we observed no effect of the manipulation on true self judgments or attributions for sadness. Second, we previously found that the unfittingness manipulation reduced true self judgments for love but did not affect attributions of love. In this study, we found that the unfittingness manipulation reduced both. Hence, in this study, the results for happiness, sadness, and love were perfectly consistent with the true self hypothesis. The only remaining inconsistency is that, for hatred, the unfittingness manipulation increased true self judgments but did not affect attributions.

The third difference between the results of Studies 2a-b emerged in the participant-level analyses. We previously found that fittingness judgments were not associated with emotion attributions after controlling for experimental condition and true self judgments. In this study, we found that fittingness judgments were significant, suggesting that they may play some independent role in shaping emotion attributions. Yet, if that's the case, then this role would seem to be very small compared with the role of true self judgments. Fittingness judgments explained only a trivial share of the variance in emotion attributions—about an eighth of the proportion explained by true self judgments.

In sum, the results of these first two studies provide strong evidence against the fittingness hypothesis and some evidence in favor of the true self hypothesis. In the next study we sought to test the true self hypothesis more directly.

7. Study 3

This study took the inverse approach from Studies 2a-b. Instead of manipulating the fittingness of the agents' emotions, we manipulated whether the agents' emotions reflected their true selves. That is, we described each agent's feelings either as arising from "deep down inside" or as being merely "on the surface." The true self hypothesis predicts that this will affect emotion attributions across the board. By contrast, the fittingness hypothesis would not predict any effect on emotion attributions, except in cases where the manipulation also happens to influence fittingness judgments. Although the manipulation did not explicitly provide any information about whether the agent's emotions were fitting, we expected that, as in Studies 2a-b, participants may make certain inferences about the emotions' fittingness. If so, then, the fittingness hypothesis would predict an effect of the manipulation on emotion attributions—but only in these specific cases, and not across the board.

7.1. Method

Participants. We recruited 604 participants from Prolific. As preregistered, we excluded participants ($n = 9$) who failed the quality-check question. This left $N = 595$ participants ($M_{age} = 38.08$, $SD_{age} = 13.80$; 50.4% identified as men, 48.6% women, and 1.0% other gender or prefer not to say; 7.9% identified as Asian, 6.6% Black or African American, 6.4% Hispanic or Latinx/Latiné, 74.4% White, 4.2% mixed race, < 1% other race or prefer not to say).

Procedure and Measures. Participants read short vignettes featuring an agent named Mario. We randomized participants to one of eight vignettes in a 2 (true self: true, not true) x 4 (emotion: happiness, sadness, love, hatred) experimental design. In the happiness, sadness, love, and hatred conditions, Mario was described, respectively, as: satisfied with life and in a good mood most of the time; dissatisfied with life and in a bad mood most of the time; having very romantic feelings for a woman named Dax; and as having very antagonistic feelings towards a woman named Dax.

The last paragraph of each vignette described Mario's feelings as either welling up from deep inside or as being merely superficial. For example, in the happiness conditions, the vignette read: "…when Mario is alone and thinks about his life, [something / nothing] deep inside of him stirs. It is like some core part of his being [sees his life as essentially

good / doesn't see his life as essentially good or bad]. For this reason, his cheerful feelings [always / never] seem to well up from deep inside of him—they're [not just / just] on the surface." Brackets indicate text that varied depending on whether participants were in the true condition (before the slash) or not true condition (after the slash). We used directly parallel descriptions for the other emotions. The measures were the same as in Study 2a.

7.2. Results

Fig. 4 shows the means and distributions for all three variables across conditions. The key question in this study was whether the effects on emotion attributions (the third column in Fig. 4) would look more like the effects on fittingness judgments (first column) or the effects on true self judgments (second column). We ran three separate factorial ANOVAs to examine the effects of experimental condition on each dependent variable.

For fittingness judgments, there was a very large main effect of the manipulation, $F(1, 587) = 274.00$, $p < .001$, $\eta_p^2 = 0.32$ [0.26, 0.37], as well as a main effect of emotion, $F(3, 587) = 39.18$, $p < .001$, $\eta_p^2 = 0.17$ [0.11, 0.22], and an interaction, $F(3, 587) = 11.87$, $p < .001$, $\eta_p^2 = 0.06$ [0.02, 0.09]. Decomposing the interaction revealed that, although the size of the manipulation's effect varied somewhat across emotions, it was significant for each (all $ps < 0.001$). Curiously, participants' true self judgments tended to be lower for hatred, and especially sadness, compared with other emotions. Even in the True conditions, participants' true self judgments were noticeably lower for sadness ($M = 44.39$, $SD = 26.85$), compared with hatred ($M = 63.49$, $SD = 26.03$), love ($M = 77.03$, $SD = 21.58$), and happiness ($M = 84.14$, $SD = 18.24$). Nevertheless, the manipulation was successful in reducing true self judgments for all emotions, including sadness, $t(587) = 4.36$, $p < .001$, which enables us to test whether this manipulation also reduced attributions of each emotion.

For true self judgments, there were main effects of the manipulation, $F(1, 587) = 32.95$, $p < .001$, $\eta_p^2 = 0.05$ [0.02,0.09], and emotion, $F(3, 587) = 212.18$, $p < .001$, $\eta_p^2 = 0.52$ [0.47, 0.57], plus an interaction effect, $F(3, 587) = 6.43$, $p < .001$, $\eta_p^2 = 0.03$ [0.01, 0.06]. Decomposing

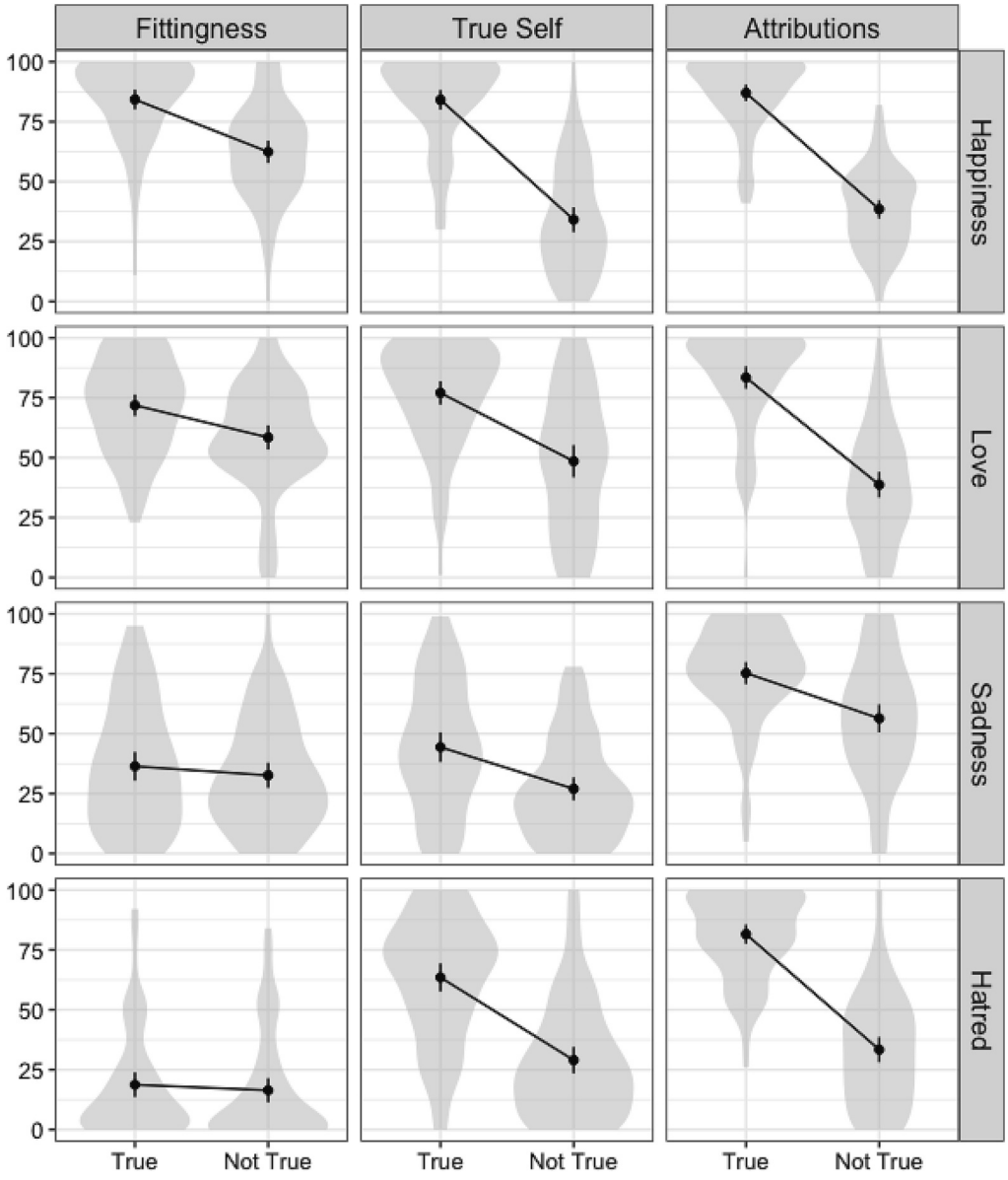

Fig. 4. Violin Plots for Dependent Variables Across Experimental Conditions in Study 3 | Points indicate condition means. Error bars indicate 95% confidence intervals.

the interaction revealed that the manipulation reduced fittingness judgments for happiness, $t(587) = 6.09$, $p < .001$, and love, $t(587) = 3.72$, $p < .001$, but not sadness, $t(587) = 1.04$, $p = .301$, or hatred, $t(587) = 0.65$, $p = .513$.

Thus, for happiness and love, the true self manipulation had parallel effects, reducing both true self judgments and fittingness judgments. Yet, for sadness and hatred, the manipulation reduced true self judgments without reducing fittingness judgments. This difference enables us to test whether the effects of the true self manipulation on emotion attributions look like the effects on true self judgments (i.e., reductions across the board) or like the effects on fittingness judgments (i.e., reductions for happiness and love but no effects for sadness and hatred).

For emotion attributions, there was a main effect of the true self manipulation, $F(1, 587) = 561.24$, $p < .001$, $\eta_p^2 = 0.49$ [0.44, 0.54], and emotion, $F(3, 587) = 4.27$, $p = .005$, $\eta_p^2 = 0.02$ [0.00, 0.05], plus an interaction, $F(3, 587) = 17.42$, $p < .001$, $\eta_p^2 = 0.08$ [0.04, 0.12]. Decomposing the interaction revealed that, although the manipulation's effect was smaller for sadness than for the other emotions, it was significant for all four (all $ps < 0.001$).

Shifting to the participant-level analyses, we used ANCOVAs to test for unique associations between these two kinds of judgments and emotion attributions. Full results of these models are shown in Table 1. In Model 1, in addition to the experimental factors and their interaction, we included a term for participants' fittingness judgments. This term proved significant. In Model 2, we added a term for participants' true self judgments. In this model, both kinds of judgments were significant, but true self judgments explained substantially more variance in emotion attributions than did fittingness judgments ($\eta_p^2 = 0.10$ and 0.02 respectively).

7.3. Discussion

The true self manipulation led to large reductions in true self judgments across the board, but only reduced fittingness judgments for two emotions (happiness and love). The effects on attributions then mirrored the effects on true self judgments—large reductions across the board. These results are exactly what the true self hypothesis would predict, but it's unclear how the fittingness hypothesis could explain them.

The results of the participant-level analyses mirrored those of Study 2b. That is, we found that fittingness judgments and true self judgments each significantly predicted emotion attributions after controlling for experimental condition. Yet, as before, fittingness judgments explained only a minute fraction of the variance, whereas true self judgments explained substantially more. This result suggests that there may be some role for fittingness judgments in shaping emotion attributions. Yet this role would appear to be quite small.

In sum, across Studies 2a-3, we found that an unfittingness manipulation reduced attributions only of emotions for which it also reduced true self judgments, whereas a true self manipulation reduced attributions of all emotions, including those for which it did not reduce fittingness judgments. Overall, then, our findings are clearly favoring the true self hypothesis and not the fittingness hypothesis.

8. Study 4

In the introduction, we speculated that the effect of evaluative judgments on emotion attributions might arise when people think that someone is experiencing an emotion only in a limited, mundane sense, but not in a more complete and profound sense denoted by phrases like "true love" and "true happiness." If this is right, then the emotion attribution questions we used in the previous studies were ambiguous, and we would observe larger effects if we asked specifically about "true" emotions. After all, the statement that a person is experiencing a particular emotion could be interpreted as the statement that the person is experiencing the emotion in the mundane sense, or as the statement that they are experiencing it in the profound sense. If we were to ask participants specifically whether the agents are experiencing *true* happiness, love, sadness, and hatred, then there would be no ambiguity, and our experimental manipulation should have a more substantial effect on emotion attributions.

Accordingly, in this study, used the same materials as in Study 2a, except that instead of assessing emotion attributions in the typical fashion, we asked participants whether the agents experienced *true* happiness, love, sadness, and hatred.

8.1. Method

Due to an oversight, this study was not pre-registered.

Participants. We recruited 603 participants from Prolific, an online research platform. As in Study 2, we excluded participants ($n = 4$) who did not correctly answer the quality-check question at the end of the survey. This left $N = 599$ participants ($M_{age} = 38.36$, $SD_{age} = 14.32$; 49.4% identified as men, 48.6% women, and 2.0% other gender or prefer not to say; 8.5% identified as Asian, 4.3% Black or African American, 5.2% Hispanic or Latinx/Latiné, 74.4% White, 5.8% mixed race, < 1% other race, < 1% prefer not to say).

Procedure and Measures. The procedure and measures for this study were identical to the ones used in Study 2a, with one exception. Instead of the regular emotion attribution question, we asked participants, "Is what Mario feels *true [emotion]?*"

8.2. Results

Fig. 5 shows the means and distributions for all three variables across conditions. Since the phrasing of the fittingness and true self questions were identical to Study 2a, we expected the results for those variables to be the same as before. The key question in these analyses is therefore whether the results for true emotion attributions looked different from the results for regular emotion attributions from Study 2a. We ran three separate factorial ANOVAs to examine the effects of experimental condition on each dependent variable.

For fittingness judgments, there was a very large main effect of the manipulation, $F(1, 593) = 554.37$, $p < .001$, $\eta_p^2 = 0.48$ [0.43, 0.53], as well as an effect of emotion, $F(3, 593) = 19.50$, $p < .001$, $\eta_p^2 = 0.09$ [0.05, 0.13], and an interaction, $F(3, 593) = 10.19$, $p < .001$, $\eta_p^2 = 0.05$ [0.02, 0.08]. Decomposing the interaction revealed that, although the effect of the manipulation was somewhat smaller for sadness, it was significant in every case (all $ps < 0.001$). Hence, as expected and as we found in Study 2a, the unfittingness manipulation successfully reduced fittingness judgments for all emotions.

For true self judgments, there was a main effect of the manipulation, $F(1, 593) = 16.52$, $p < .001$, $\eta_p^2 = 0.03$ [0.01, 0.06], but not emotion, $F(3, 593) = 1.19$, $p = .325$, $\eta_p^2 = 0.01$ [0.00, 0.02], though there was an interaction, $F(3, 593) = 36.25$, $p < .001$, $\eta_p^2 = 0.16$ [0.10, 0.21]. Decomposing the interaction revealed that the unfittingness manipulation reduced true self judgments for happiness, $t(591) = -9.53$, $p < .001$, and love, $t(591) = -4.11$, $p < .001$, had no effect for sadness, $t(591) = 1.74$, $p = .082$, and increased true self judgments for hatred, $t(591) = 3.79$, $p < .001$. Thus, these results are identical to those of Study 2b, but different from those of Study 2a, where the manipulation reduced true self judgments for sadness.

For true emotion attributions, there was a main effect of the unfittingness manipulation, $F(1, 593) = 90.66$, $p < .001$, $\eta_p^2 = 0.13$ [0.09, 0.18], as well as an effect of emotion, $F(3, 593) = 13.68$, $p < .001$, $\eta_p^2 = 0.06$ [0.03, 0.10], and an interaction, $F(3, 593) = 32.62$, $p < .001$, $\eta_p^2 = 0.14$ [0.09, 0.19]. Decomposing the interaction revealed that the unfittingness manipulation reduced true emotion attributions for happiness, $t(591) = -12.21$, $p < .001$, $d = 2.00$, 95% CI: [1.66, 2.35], and love, $t(591) = -5.94$, $p < .001$, $d = 0.97$, 95% CI: [0.64, 1.29].

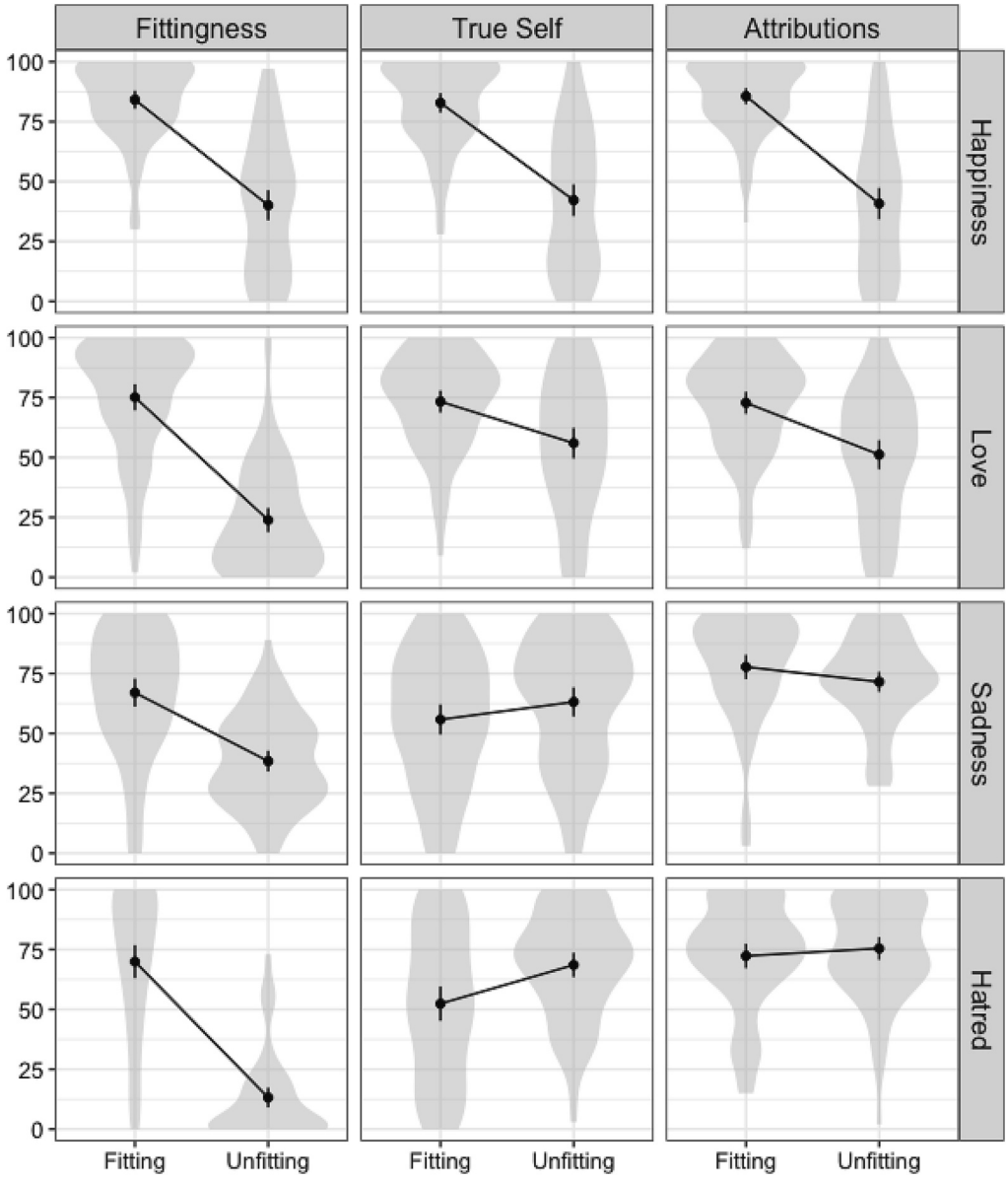

Fig. 5. Violin Plots for Dependent Variables Across Experimental Conditions in Study 4 | Points indicate condition means. Error bars indicate 95% confidence intervals.

However, it had no effect for sadness, $t(591) = -1.72$, $p = .087$, $d = 0.28$, 95% CI: [−0.04, 0.60], or hatred, $t(591) = 0.85$, $p = .396$, $d = -0.14$, 95% CI:[−0.46, 0.18]. Thus, the overall pattern of results is identical to what we observed in Study 2b.

Shifting to the participant-level analyses, we used ANCOVAs to test for unique associations between these two kinds of judgments and emotion attributions. Full results are shown in Table 1. The overall pattern of results was much the same as in Study 2b, only the variance explained by true self judgments was increased. In Model 1, true self judgments accounted for about a quarter of the variance in true emotion attributions ($\eta_p^2 = 0.22$). In Model 2, fittingness judgments also significantly predicted true emotion attributions. However, they explained only a small proportion of the variance ($\eta_p^2 = 0.03$) and adding them to the model did not appreciably reduce the variance explained by true self judgments.

8.3. Discussion

The overall pattern of results in this study are very similar to those of Studies 2a-b. However, there is one important difference, reflecting the change we made to the emotion attribution question. In the cases where the experimental manipulation affected attributions (i.e., happiness and love), it had larger effects than we previously observed. For example, in this study, the size of the effect on happiness attributions ($d = 2.00$) was almost exactly double what we observed in Study 2a ($d = 1.02$). This supports our speculation that these effects on emotion attributions emerge because people think that the agent's emotions aren't "true." To illustrate, if a person is in a good mood and satisfied with life but is living a bad life, people generally agree that the person is happy, though they are less inclined to agree completely. However, when asked whether such a person is experiencing *true* happiness, people tend to be more ambivalent, and many people disagree.

As in Studies 2a-b, the fittingness hypothesis did a poor job of explaining the results. The unfittingness manipulation reduced fittingness judgments across the board, but it only reduced attributions of true happiness and true love, not true sadness or true hatred. In the participant-level analyses, fittingness judgments significantly predicted true emotion attributions while controlling for experimental condition

and true self judgments. Yet, as before, they only explained a trivial share of the variance.

By contrast, the true self hypothesis continued to do a better job of explaining the results. In both cases where the unfittingness manipulation reduced true emotion attributions (happiness and love), it also reduced true self judgments. In both cases where it did *not* reduce true emotion attributions (sadness and hatred), it did not reduce true self judgments. Additionally, in participant-level analyses, true self judgments accounted for a substantial proportion of the variance in true emotion attributions. However, the true self hypothesis again faces a problem that we observed previously. Because the experimental manipulation *increased* true self judgments for hatred, the true self hypothesis would predict an increase in attributions of true hatred, which we did not observe.

9. Study 5

This study used the same materials as in Study 3 except that, instead of assessing regular emotion attributions, we assessed *true* emotion attributions. That is, we manipulated whether an agent's emotion was described as reflecting their true self or as being more superficial, and tested for effects on attributions of true happiness, love, sadness, and hatred

10. Method

Participants. We recruited 603 participants from Prolific. As preregistered, we excluded participants ($n = 3$) who did not respond to a quality-check question ("What day of the week is it?") at the end of the

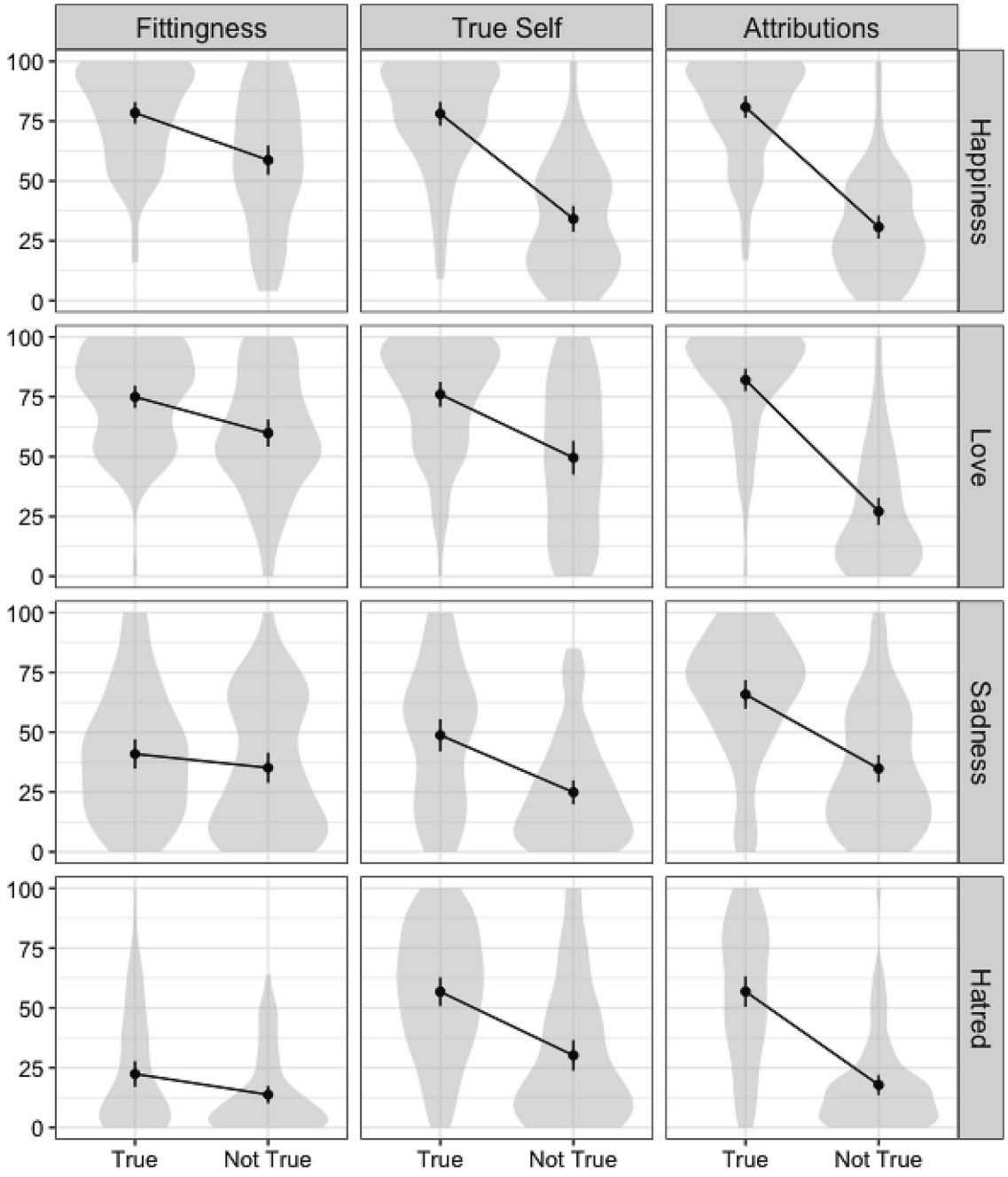

Fig. 6. Violin Plots for Dependent Variables Across Experimental Conditions in Study 5 | Points indicate condition means. Error bars indicate 95% confidence intervals.

survey. This left $N = 600$ participants ($M_{age} = 36.28$, $SD_{age} = 13.56$; 49.8% identified as men, 48.8% women, and 1.3% other gender or prefer not to say; 12.5% identified as Asian, 7.2% Black or African American, 6.3% Hispanic or Latinx/Latiné, 65.2% White, 7.0% mixed race, < 1% other race, < 1% prefer not to say).

Procedure and Measures. The procedure and measures were identical to those used in Study 3, with one exception. Instead of the emotion attribution question, we used the true emotion attribution question from Study 4.

10.1. Results

Fig. 6 shows the means and distributions for all three variables across conditions. Since the phrasing of the fittingness and true self questions were identical to Study 3, we expected the results for those variables to be the same as before. The key question in these analyses is therefore whether the results for true emotion attributions looked different from the results for regular emotion attributions from Study 3. We ran three separate factorial ANOVAs to examine the effects of experimental condition on each dependent variable.

For true self judgments, there was a very large main effect of the manipulation, $F(1, 592) = 202.78$, $p < .001$, $\eta_p^2 = 0.26$ [0.20, 0.31], as well as an effect of emotion, $F(3, 592) = 30.78$, $p < .001$, $\eta_p^2 = 0.13$ [0.09, 0.18], and an interaction, $F(3, 592) = 4.75$, $p = .003$, $\eta_p^2 = 0.02$ [0.02, 0.05]. Decomposing the interaction revealed that, although the size of the manipulation's effect varied somewhat across emotions, it was significant for each (all $ps < 0.001$). Hence, as expected, and as we found in Study 3, the true self manipulation successfully reduced true self judgments for all emotions.

For fittingness judgments, there were main effects of the manipulation, $F(1, 592) = 40.44$, $p < .001$, $\eta_p^2 = 0.06$ [0.03, 0.11], and emotion, $F(3, 592) = 160.26$, $p < .001$, $\eta_p^2 = 0.45$ [0.39, 0.50], as well as an interaction effect, $F(3, 592) = 2.64$, $p < .001$, $\eta_p^2 = 0.01$ [0.00, 0.03]. Decomposing the interaction revealed that the manipulation reduced fittingness judgments for happiness, $t(592) = -5.10$, $p < .001$, love, $t(592) = -3.92$, $p < .001$, and hatred, $t(592) = -2.24$, $p = .025$, but not sadness, $t(592) = -1.48$, $p = .140$. Thus, these results are the same as those of Study 3, except that, previously, we did not observe the reduction in fittingness judgments for hatred.

For true emotion attributions, there was a main effect of the true self manipulation, $F(1, 592) = 524.66$, $p < .001$, $\eta_p^2 = 0.47$ [0.42, 0.52], an effect of emotion, $F(3, 592) = 19.62$, $p < .001$, $\eta_p^2 = 0.09$ [0.05, 0.13], and an interaction, $F(3, 592) = 8.01$, $p < .001$, $\eta_p^2 = 0.04$ [0.01, 0.07]. Decomposing the interaction revealed that, although the size of the manipulation's effect varied somewhat across emotions, it significantly reduced true emotion attributions in every case (all $ps < 0.001$). Thus, these results are identical to those of Study 3.

Shifting to the participant-level analyses, we used ANCOVAs to test for unique associations between these two kinds of judgments and emotion attributions. Full results of these models are shown in Table 1. In Model 1, fittingness judgments significantly predicted true emotion attributions and explained a substantial proportion of the variance ($\eta_p^2 = 0.13$). In Model 2, true self judgments also emerged as significant and explained a similar proportion of the variance in true emotion attributions ($\eta_p^2 = 0.16$).[1] Strikingly, adding true self judgments to the model substantially reduced the variance explained by fittingness judgments ($\eta_p^2 = 0.07$).

[1] Although Model 1 was pre-registered, Model 2 was added post hoc for consistency with the other studies.

10.2. Discussion

The results of this study are nearly identical to those of Study 3. The true self manipulation led to large reductions in true self judgments but only small reductions in fittingness judgments and only for some emotions. The effects on "true" emotion attributions then mirrored the effects on true self judgments—large reductions across the board. Hence, the results do not align with the predictions of the fittingness hypothesis, but they are precisely in line with the predictions of the true self hypothesis.

In Study 4 we found that manipulating the fittingness of these four emotions only affected whether people thought that happiness and love were "true." It had no effect on whether people thought that sadness or hatred were "true." By contrast, in this study, we found that manipulating whether these emotions reflect the agent's true self affected judgments about the trueness of all four emotions. Indeed, the effects of a true self manipulation on true emotion attributions were larger than the effects observed on regular emotion attributions, continuing to support the idea that the effects on regular emotion attributions emerged because people thought that the agent's emotions weren't "true." Hence, these results suggest that people think a person's emotion is "true" when they think that the emotion reflects that person's true self.

11. General discussion

In six experiments, we investigated two hypotheses about how observers attribute emotions to others. Study 1 found that people think a wide range of different emotions, both positive and negative, can be fitting, reflect a person's true self, and be called "true." In Studies 2a-b, we presented participants with vignettes and experimentally manipulated whether an agent's happiness, love, sadness, and hatred were fitting or unfitting. In Study 2a, we used vignettes that gave concrete details about the agent's specific life circumstances. In Study 2b, we used highly abstract vignettes that did not provide any concrete details and only described the abstract structure of the case. We observed very similar results in each study. Contrary to the fittingness hypothesis, the manipulations only affected attributions of happiness and love, having no effect on the attributions of sadness or hatred. In Study 3 we manipulated whether an agent's happiness, love, sadness, and hatred reflected the agent's true self. In line with the true self hypothesis, this manipulation influenced attributions of all four emotions.

In Studies 4–5 we modified the dependent variable slightly, assessing attributions of "true happiness," "true love," "true sadness," and "true hatred." The results looked quite similar to those of Studies 2a-b and 3. The unfittingness manipulation influenced attributions of true happiness and true love, but not true sadness or true hatred, whereas the true self manipulation affected true emotion attributions across the board. In these two studies, when we observed effects on attributions, these effects were larger than in the previous studies. We interpret this as indicating that, when the unfittingness and true self manipulations affected regular emotion attributions, this was because these manipulations led people to think that the agents' emotions were not "true."

In Studies 2–5, we also examined the associations between participants' own judgments. Controlling for experimental condition, participants' fittingness judgments significantly predicted their emotion attributions in four out of five studies. However, fittingness judgments consistently explained only a very small amount of variance in emotion attributions. By contrast, participants' judgments about whether the emotion reflected the agent's true self significantly predicted emotion attributions in all five studies and explained substantially more of the variance in emotion attributions.

11.1. Complex links between fittingness and true self judgments

Across five experiments, we found that the unfittingness manipulations consistently influenced true self judgments, and the true self

manipulations consistently influenced fittingness judgments. Moreover, these effects revealed a complex interaction pattern. For example, whereas the unfittingness manipulation led to small reductions in true self judgments for happiness and love, it led to a small *increase* in true self judgments for hatred (and sadness in Study 2a).

What might explain these results? As discussed in the introduction, past research has found that people tend to assume that the true self calls one to be good. This could explain why fitting happiness and love are assumed to reflect a person's true self. However, there is also evidence that, under certain conditions, people think that morally bad actions and feelings can reflect a person's true self. In other words, people sometimes override their default assumption that the true self is good. Perhaps this is what is happening when people consider unfitting hatred. If the target of someone's hatred is perfectly kind and wonderful, then it seems unlikely that there would be any sort of external, social pressure to hate that target person. Hence, if a person hates them nonetheless, people may override their default assumption that the person's true self is good and conclude that the hatred reflects the person's true self. Indeed, previous research has found that people think hateful, racist behavior is more reflective of a person's true self when that person was *not* raised to be racist (Daigle & Demaree-Cotton, 2022). Hence, it may be that, in line with correspondent inference theory (Jones & Harris, 1967), when a person is not at all encouraged to feel hatred but feels hatred nonetheless, people are more inclined to think that the hatred reflects the person that they are deep down inside.

In any case, regardless of what explains the complex patterns of effects that we observed, the fact that our experimental manipulations in Studies 2–5 had different effects on fittingness judgments and true self judgments enables us to test the two hypotheses. Whenever our experimental manipulations affected fittingness judgments, the fittingness hypothesis predicts a corresponding effect (in the same direction, though not necessarily of the same size) on emotion attributions. Similarly, whenever the experimental manipulations affected true self judgments, the true self hypothesis predicts a corresponding effect on emotion attributions. The fact that the experimental manipulations led to complex patterns of effects on fittingness and true self judgments means that we can compare these patterns to the pattern for emotion attributions and determine which hypothesis' predictions received more support.

11.2. The fittingness hypothesis

The core idea of the fittingness hypothesis is that when people attribute emotions to others, these attributions depend in part on their beliefs about the agent's external circumstances. In particular, the hypothesis is that people's attributions of various emotions depend not only on whether they think that the agent has the right mental states but also on whether they think that the agent's actual situation makes those mental states *fitting*.

If one looked only at the results for attributions of happiness or love, the fittingness hypothesis might seem quite plausible. People are less willing to attribute these emotions when they consider these emotions unfitting. They are also less inclined to call these feelings "true happiness" and "true love." However, when one looks at a larger range of emotions, it becomes clear that there is strong evidence against this hypothesis. People appear to be no less willing to attribute sadness or hatred (or consider them "true") when these emotions are unfitting versus fitting. In other words, if someone appears to be sad, then people think that the person is sad, regardless of whether it actually makes sense for this person to feel sad. Similarly, if someone appears to hate another person, then people think that the person feels hatred, even if that other person doesn't merit hatred at all.

In four out of the five experimental studies, our participant-level analyses revealed that fittingness judgments significantly predict emotion attributions (after controlling for experimental condition). Yet, they explain only a trivial amount of the variance in these attributions. Furthermore, given the effects of the experimental manipulations, it's difficult to see how this association could reflect a causal relationship. Hence, the fittingness hypothesis—at least as we have interpreted it here, and as it is typically developed in the philosophical literature—no longer looks viable as a claim about ordinary thinking.

Future work could continue to explore hypotheses that involve the notion of fittingness. But, to do so, this research would have to develop new theoretical frameworks that do not mistakenly predict an effect of fittingness for cases like the ones involving sadness and hatred in the present studies. For example, future theorizing about fittingness and emotion concepts could perhaps lead to a restricted version of the fittingness hypothesis that explains why fittingness affects attributions of only some emotions. If future theoretical work does lead to the development of such a hypothesis, it might be fruitfully investigated in empirical research.

11.3. The true self hypothesis

The core idea behind the true self hypothesis is that people distinguish between how a person feels "on the surface" and how they feel "deep down" in their true self. Thus, when observers attribute emotions, their attributions depend partly on their beliefs about whether the target person's feelings reflect the target's true self.

Overall, our results provide strong support for this hypothesis. First, in Studies 2a-b and 4, the true self hypothesis accurately predicts when the unfittingness manipulation would decrease emotion attributions—namely, in those cases where it also decreases true self judgments. We found that, although the unfittingness manipulation did not say anything about the agents' true selves, it led to decreases in true self judgments for happiness and love but not for sadness or hatred. Crucially, we also only observed decreases in attributions of happiness and love, and not sadness or hatred. Second, the true self hypothesis received strong support in Studies 3 and 5, where we manipulated whether the agents' emotions were described as reflecting his true self. In these studies, we observed decreases in attributions of all emotions (including those for which the manipulation did *not* decrease fittingness judgments). Third, in all four studies, true self judgments explained more variance in emotion attributions than did fittingness judgments, and even more than the unfittingness manipulations in Studies 2a-b and 4. Given all this, it seems that there is strong reason to think that the true self hypothesis is getting at something important about how people attribute emotions to others.

Yet, the true self hypothesis faces a problem when it comes to one specific finding. We did not merely find that the unfittingness manipulations did not *decrease* true self judgments for sadness and hatred. Instead, we found an effect in the opposite direction. The unfittingness manipulation *increased* true self judgments for hatred in Studies 2a-b and 4, as well as for sadness (though only in Study 2a). That is, when we described the agent as hating someone perfectly kind and decent, participants thought that the agent's hatred was more reflective of the agent's true self than when we described the agent as hating someone horrible and vile. This poses a problem for the true self hypothesis insofar as we did not find a corresponding effect on attributions of hatred. For hatred, and perhaps sadness as well, when people are told that the agent's situation does not merit the emotion, people don't show any reduction in their tendency to attribute those emotions, but they also don't show any increase.

What does this finding show about the true self hypothesis? One possible view would be that, despite the apparent evidence in its favor, this finding shows that the hypothesis is fundamentally mistaken. Another possible view would be that, given all of the other evidence in favor of the hypothesis, we have strong reason to suspect that the true self hypothesis is basically on the right track. That is, *some version* of the true self hypothesis is probably correct, just not the version that we have focused on here. This second view implies that we need to reflect further on the precise way in which intuitions about the true self influence

emotion attributions.

The version of the true self hypothesis that we tested in these studies claims that emotion attributions are influenced by the degree to which the emotion *reflects* a person's true self. Yet, previous research on the true self has assessed a range of subtly, but perhaps importantly, different kinds of intuitions about the true self. These have included: whether something is *caused* by an agent's true self (De Freitas et al., 2018); whether an agent is *drawn towards* something by their true self (B. Phillips, 2022); or whether something *goes against* an agent's true self (Newman et al., 2015). Although these judgments are likely closely related, they may also come apart.

For instance, even if an emotion does not reflect one's true self, it would not follow that one's true self is *opposed* to that emotion. Prior research on the effect of moral judgments on happiness attributions suggests that the effect arises because, when an agent acts immorally, people typically think that the agent is internally conflicted (Prinzing & Fredrickson, 2022). Hence, it could be that the true self hypothesis would fare better if it were reformulated in terms of judgments about a (lack of) conflict or opposition between the agent's surface-level feelings and their true self. We tested this possibility in another experiment, reported in the Supplemental Materials. We copied the materials from Study 2a, but changed the true self question to ask whether deep down in his true self the agent actually feels the opposite of what he feels on the surface. This change brought the results for sadness in line with the predictions of the true self hypothesis (as we observed in Studies 2b and 4, but not 2a). However, for hatred, we still saw an effect on true self judgments without an effect on emotion attributions. Hence, this reformulation of the true self hypothesis still did not perfectly explain the results.

Future studies could continue to explore how people conceive of the true self, and its relationship to the surface self, in order to narrow in more precisely on the kind of true self judgments that drive emotion attributions. Though, of course, such studies should use new materials and experimental designs so as not to simply design a measure to capture known results.

11.4. What are "True" Emotions?

In Studies 4–5, we switched from simply asking whether an agent is "happy," "sad," and so on to asking whether the agent is experiencing "*true* happiness," "*true* sadness," and so on. Framing the question in terms of "true" emotions did not lead to a completely different pattern of effects, but it did lead to an amplification of some of the effects observed in the earlier studies. For example, Study 2a found that when participants were simply asked whether the agent was sad or happy, there was no effect of the unfittingness manipulation for sadness, but there was an effect for happiness. When we switch to "true sadness" and "true happiness," we still get no effect for sadness, but the effect for happiness becomes considerably larger than it was without the modifier "true."

Past theoretical work on true emotions has centered on the idea that emotions are true when they are fitting (De Sousa, 2002; Hamlyn, 1989; Salmela, 2006; Solomon, 2002). Yet these results from Studies 4–5 indicate that this is not what ordinary people think. We found that the unfittingness manipulation only decreased attributions of true emotions when it also decreased true self judgments, while the true self manipulation affected true emotion attributions across the board. These findings provide at least some initial support for a very different hypothesis about what people mean when they call an emotion true—namely, that an emotion is seen as true to the extent that it is seen as related in a certain way to the agent's true self.

This hypothesis fits in with a larger picture of what people are doing when they speak of "true happiness," "true sadness," and so forth. Consider an agent who is happy at a superficial level but who is not happy deep down in her true self. If people are simply asked whether or not this agent is happy, they may feel uncertain or torn in different directions. They might be inclined to give a mixed answer like: "In a certain sense, she is happy, but in a deeper sense, she isn't happy at all." However, if they are asked whether this agent has "true happiness," they might not find the question nearly as difficult. It would be clear that they are not being asked whether the agent is happy on the surface but are instead being asked whether the agent is happy in her true self.

On this hypothesis, then, the question as to whether an agent is experiencing "*true* happiness" does not simply mean the exact same thing as the question as to whether she is "happy," but it is also not a question about a completely different thing. Rather, the seemingly straightforward question as to whether an agent is happy can be understood in slightly different ways, and the question about true happiness reflects one way of specifying the broader question.

12. Conclusion and future directions

The present studies investigated people's intuitive judgments about whether someone is experiencing a particular emotion, whether that emotion is fitting, and whether it reflects the person's true self. In future research, it might be helpful to use similar methods to explore other, closely related phenomena. In recent years, there has been a surge of research on how people ordinarily think about the good life—encompassing not only judgments about emotions like happiness, sadness, love, and hatred, but also judgments of overall well-being (Bronsteen, Leiter, Masur, & Tobia, 2022), meaning in life (Fuhrer & Cova, 2021; Prinzing, De Freitas, & Fredrickson, 2021) and purpose in life (Taylor, Kalbach, & Rose, 2019). The methods employed in the present studies could potentially be useful for future work on all of these topics. For each of these topics, we face a question about whether people's judgments are driven by judgments about whether certain mental states are fitting versus whether those mental states reflect the true self. In each case, further research could explore that question using some of the methods we have used here.

CRediT authorship contribution statement

Michael Prinzing: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Brian Earp:** Conceptualization, Writing – review & editing. **Joshua Knobe:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision.

Data availability

Data, materials, analytic code, and pre-registration forms are available on OSF: https://osf.io/fvbh6/.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2023.105579.

References

Bronsteen, J., Leiter, B., Masur, J. S., & Tobia, K. (2022). The folk theory of well-being. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4146762

Chen, X., Harris, P. L., & Yang, F. (2023). Beyond enjoyment: Young children consider the normative goodness of activity engagement when attributing happiness. *Journal of Experimental Child Psychology, 228*, Article 105608. https://doi.org/10.1016/j.jecp.2022.105608

Christy, A. G., Schlegel, R. J., & Cimpian, A. (2019). Why do people believe in a "true self"? The role of essentialist reasoning about personal identity and the self. *Journal of Personality and Social Psychology, 117*(2), 386. https://doi.org/10.31234/osf.io/k3jba

Daigle, J. L., & Demaree-Cotton, J. (2022). Blame mitigation: A less tidy take and its philosophical implications. *Philosophical Psychology, 35*(4), 490–521. https://doi.org/10.1080/09515089.2021.2000594

D'Arms, J., & Jacobson, D. (2000). The moralistic fallacy: On the "appropriateness" of emotions. *Philosophical and Phenomenological Research, 61*(1), 65–90. https://doi.org/10.2307/2653403

De Freitas, J., & Cikara, M. (2018). Deep down my enemy is good: Thinking about the true self reduces intergroup bias. *Journal of Experimental Social Psychology, 74*, 307–316. https://doi.org/10.1016/j.jesp.2017.10.006

De Freitas, J., Cikara, M., Grossmann, I., & Schlegel, R. (2017). Origins of the belief in good true selves. *Trends in Cognitive Sciences, 21*(9), 634–636. https://doi.org/10.1016/j.tics.2017.05.009

De Freitas, J., Sarkissian, H., Newman, G. E., Grossmann, I., Brigard, F. D., Luco, A., & Knobe, J. (2018). Consistent belief in a good true self in misanthropes and three interdependent cultures. *Cognitive Science, 42*(S1), 134–160. https://doi.org/10.1111/cogs.12505

De Sousa, R. (2002). Emotional truth. *Aristotelian Society Supplementary, 76*(1), 247–263. https://doi.org/10.1111/1467-8349.00098

Díaz, R., & Reuter, K. (2020). Feeling the right way: Normative influences on people's use of emotion concepts. *Mind & Language*. https://doi.org/10.1111/mila.12279

Earp, B. D., Do, D., & Knobe, J. (2021). The ordinary concept of true love. In C. Grau, & A. Smuts (Eds.), *The Oxford handbook of philosophy of Love*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199395729.013.38.

Earp, B. D., Hannikainen, I., Dale, S., & Latham, S. (2023). Experimental philosophical bioethics, advance directives, and the true self in dementia. In A. De Block, & K. Hens (Eds.), *Experimental philosophy of medicine*. Bloomsbury.

Earp, B. D., Skorburg, J. A., Everett, J. A. C., & Savulescu, J. (2019). Addiction, identity, morality. *AJOB empirical. Bioethics, 10*(2), 136–153. https://doi.org/10.1080/23294515.2019.1590480

Fuhrer, J., & Cova, F. (2021). *What makes a life meaningful? Folk intuitions about the content and shape of meaningful lives*. https://doi.org/10.31234/osf.io/7vbh5

Hamlyn, D. W. (1989). False emotions. *Proceedings of the Aristotelian Society, Supplementary Volumes, 63*, 275–295.

Heiphetz, L., Strohminger, N., Gelman, S. A., & Young, L. L. (2018). Who am I? The role of moral beliefs in children's and adults' understanding of identity. *Journal of Experimental Social Psychology, 78*, 210–219.

Hicks, J. A., Schlegel, R. J., & Newman, G. E. (2019). Introduction to the special issue: Authenticity: Novel insights into a valued, yet elusive, concept. *Review of General Psychology, 23*(1), 3–7. https://doi.org/10.1177/1089268019829474

Hitlin, S. (2003). Values as the core of personal identity: Drawing links between two theories of self. *Social Psychology Quarterly, 66*(2), 118–137. https://doi.org/10.2307/1519843

Howard, C. (2018). Fittingness. *Philosophy Compass, 13*(11), Article e12542. https://doi.org/10.1111/phc3.12542

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology, 3*(1), 1–24. https://doi.org/10.1016/0022-1031(67)90034-0

Kim, J., Christy, A. G., Rivera, G. N., Schlegel, R. J., & Hicks, J. A. (2018). Following one's true self and the sacredness of cultural values. *Journal of Experimental Social Psychology, 76*, 100–103. https://doi.org/10.1016/j.jesp.2018.01.001

Kneer, M., & Haybron, D. (2023). Taking the morality out of happiness. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4350806

Lantian, A., Boudesseul, J., & Cova, F. (2023). Prescription for Love: an experimental investigation of Laypeople's moral disapproval of Love drugs [preprint]. *PsyArXiv.*. https://doi.org/10.31234/osf.io/572wa

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). *Package "emmeans"* [computer software]. https://cran.microsoft.com/snapshot/2018-01-13/web/packages/emmeans/emmeans.pdf.

Lopez, I. J. (1974). A fine observation on the concept of depression. In *Depression in everyday practice* (pp. 33–35). H. Huber.

Maglio, S. J., & Reich, T. (2019). Feeling certain: Gut choice, the true self, and attitude certainty. *Emotion, 19*(5), 876–888. https://doi.org/10.1037/emo0000490

Martin, J. W., Charles, S., & Heiphetz, L. (2022). Essentialist views of criminal behavior predict increased punitiveness. In J. Musolino, J. Sommer, & P. Hemmer (Eds.), *The cognitive science of belief* (1st ed., pp. 254–276). Cambridge University Press. https://doi.org/10.1017/9781009001021.019.

Martin, M. W. (2012). *Happiness and the good life*. USA: Oxford University Press.

Morgan, C. R., & Averill, J. R. (1992). True feelings, the self, and authenticity: a psychosocial perspective. In D. Franks, & V. Gecas (Eds.), *Social perspectives on emotion* (pp. 95–124). JAI Press.

Newman, G. E., Bloom, P., & Knobe, J. (2014). Value judgments and the true self. *Personality and Social Psychology Bulletin, 40*(2), 203–216. https://doi.org/10.1177/0146167213508791

Newman, G. E., Freitas, J. D., & Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science, 39*(1), 96–125. https://doi.org/10.1111/cogs.12134

Newman, G. E., Lockhart, K. L., & Keil, F. C. (2010). "End-of-life" biases in moral evaluations of others. *Cognition, 115*(2), 343–349. https://doi.org/10.1016/j.cognition.2009.12.014

Nozick, R. (1989). *The examined life*. Simon & Schuster.

Oktar, K., & Lombrozo, T. (2022). Deciding to be authentic: Intuition is favored over deliberation when authenticity matters. *Cognition, 223*, Article 105021. https://doi.org/10.1016/j.cognition.2022.105021

Phillips, B. (2022). *The true empathetic self [unpublished manuscript]*.

Phillips, J., De Freitas, J., Mott, C., Gruber, J., & Knobe, J. (2017). True happiness: The role of morality in the folk concept of happiness. *Journal of Experimental Psychology: General, 146*(2), 165–181. https://doi.org/10.1037/xge0000252

Phillips, J., Misenheimer, L., & Knobe, J. (2011). The ordinary concept of happiness (and others like it). *Emotion Review, 3*(3), 320–322. https://doi.org/10.1177/1754073911402385

Prinzing, M. M., De Freitas, J., & Fredrickson, B. L. (2021). The ordinary concept of a meaningful life: The role of subjective and objective factors in attributions of meaning. *The Journal of Positive Psychology.*. https://doi.org/10.1080/17439760.2021.1897866

Prinzing, M. M., & Fredrickson, B. L. (2022). *No peace for the wicked? Immorality is (usually) thought to disrupt intrapersonal harmony*. Unpublished Manuscript.

Salmela, M. (2006). True emotions. *The Philosophical Quarterly, 56*(224), 382–405. https://doi.org/10.1111/j.1467-9213.2006.00448.x

Solomon, R. C. (2002). *True to our feelings: What our emotions are really telling us*. Oxford University Press.

Strohminger, N., Knobe, J., & Newman, G. (2017). The true self: A psychological concept distinct from the self. *Perspectives on Psychological Science, 12*(4), 551–560. https://doi.org/10.1177/1745691616689495

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition, 131*(1), 159–171. https://doi.org/10.1016/j.cognition.2013.12.005

Tatarkiewicz, W. (1976). *Analysis of happiness*. Martinus Nijhoff Publishers.

Taylor, M., Kalbach, C., & Rose, D. (2019). *Teleology and personal identity*. Unpublished Manuscript.

Warburton, N. (2011). *A little history of philosophy*. Yale University Press. https://doi.org/10.12987/9780300177541

Yang, F., Knobe, J., & Dunham, Y. (2020). Happiness is from the soul: The nature and origins of our happiness concept. *Journal of Experimental Psychology. General*. https://doi.org/10.1037/xge0000790

Zhang, Y., & Alicke, M. (2021). My true self is better than yours: Comparative Bias in true self judgments. *Personality and Social Psychology Bulletin, 47*(2), 216–231. https://doi.org/10.1177/0146167220919213