Self-verification and the perceived reliability of uncertain feedback sources

Douglas B. Markant (dmarkant@charlotte.edu)
Department of Psychological Science
University of North Carolina at Charlotte
9201 University City Blvd., Charlotte, NC 28223 USA

Abstract

People often have a preference for "self-verifying" feedback that confirms their existing self-views. Self-verification can reinforce existing self-views and prevent opportunities to learn from alternative perspectives, as when people with low self-esteem prefer feedback that validates negative self-beliefs. Past work suggests that a major driver of self-verification is a desire for accurate self-assessment, where disconfirmatory feedback that contradicts existing self-views creates doubt about the credibility of the feedback source. The aim of this study was to develop a formal account of self-verification based on a Bayesian model of source reliability. Findings from a behavioral experiment aligned with the model's prediction that confirmatory feedback about traits central to one's self-concept enhances the perceived reliability of a source, while disconfirmatory feedback leads to lower reliability and disinterest in further feedback. This approach clarifies why seemingly biased feedback seeking behaviors may be motivated by rational epistemic concerns about source credibility.

Keywords: self-verification, confirmation bias, feedback seeking

Feedback from others plays a central role in shaping how people view themselves. Whether the source of feedback is a romantic partner, a teacher, or a therapist, hearing how one is perceived by someone else can be informative and potentially change one's self-views. Feedback-seeking behaviors may accordingly be motivated by the desire for *self-assessment*: to reduce uncertainty about one's own traits, abilities, or standing in the social environment (Anseel, Lievens, & Levy, 2007). Yet there is substantial evidence that, rather than seek out feedback that might be informative or surprising, people often prefer *self-verifying* feedback that confirms their existing self-views (Swann, Rentfrow, & Sellers, 2012).

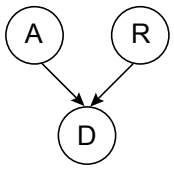
Like other forms of confirmation bias (Nickerson, 1998), self-verification can produce an echo chamber that magnifies existing self-beliefs. For instance, when given a choice between interacting with partners who have favorable or unfavorable impressions of them, many people choose to interact with the person whose impression matches their own self-view, even when that self-view is negative (Swann, Wenzlaff, Krull, & Pelham, 1992). This means that individuals with low self-esteem often prefer social feedback that is likely to reinforce their negative self-beliefs, despite the availability of other sources of positive, disconfirming feedback. Biased feedback seeking may play a role in cementing existing self-views and lead to fewer opportunities to learn from other perspectives about one's own abilities or traits.

On its face, self-verification seems inconsistent with the basic notion that people seek feedback that reduces uncertainty about their personal attributes, as is suggested by normative theories of information search in other domains (Nelson, 2005). However, past work suggests that self-verification is in fact tied to reasonable concerns about the informativeness of feedback about personal characteristics from outside sources (Swann, Stein-Seroussi, & Giesler, 1992). Many self-beliefs are formed over long periods and are informed by a rich history of personal experiences. Some theories of conceptual self-knowledge posit that self-beliefs are organized in structured, associative networks (Elder, Cheung, Davis, & Hughes, 2023). In this view, beliefs about certain traits (e.g., *"I am respectful"*) have central positions within an individual's self-concept, and as a result tend to be held with a high degree of confidence, have a strong influence on self-judgments, and are important to maintaining the coherence of their self-views. When another person offers feedback that challenges such a core belief, the recipient may be more likely to question the other person's ability to provide informative feedback rather than change the underlying belief about themselves.

A recent study illustrates the importance of perceptions of source reliability in the preference for self-verifying feedback (Szumowska, Wójcik, Szwed, & Kruglanski, 2022). When feedback was framed as coming from a non-expert peer, there was an overall preference for confirmatory, self-verifying feedback (whether positive or negative). In contrast, when feedback was described as originating from an expert, people were more willing to view disconfirmatory feedback that they expected to contradict their self-views. Self-verifying feedback choices can't be explained simply as a manifestation of confirmation bias or a desire for affirmation. People may avoid disconfirming feedback if they decide that a feedback source is not credible and, as a result, such feedback is likely to be uninformative.

Existing work thus suggests that self-verification is connected to epistemic concerns about the veracity of feedback in the presence of uncertainty about both one's own attributes and the credibility of a feedback source. While there is a rich existing literature examining the motives and consequences of self-verification, there are no formal accounts of self-verification and its relationship to the perceived credibility of information sources. In this paper I first present a for-

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*



Attribute value: A	Source reliability: R	Feedback likelihood: p(D = 1 | A, R)
1	1	1
0	1	0
1	0	.5
0	0	.5

Figure 1: The model assumes observed feedback (D) is causally related to both whether one has the attribute (A) and whether the feedback source is reliable (R). When the source is reliable, feedback matches the truth value of the trait (top 2 rows of table). When the source is unreliable, feedback is assumed to be generated at random (bottom 2 rows of table).

mal account of the effects of self-verifying feedback, building on earlier work on Bayesian models of source reliability (Pilgrim, Sanborn, Malthouse, & Hills, 2024; Merdes, Von Sydow, & Hahn, 2021; Hahn, Merdes, & Von Sydow, 2018; Jern, Chang, & Kemp, 2014). The main idea embodied by the model is that when people receive initial feedback from an external source, they simultaneously update beliefs about their own attributes and the source's reliability. Early self-verifying feedback that confirms existing self-beliefs enhances the perceived reliability of the source, while early disconfirmatory feedback leads to lower perceived reliability. Perceived reliability in turn affects whether people expect further feedback from the same source to be informative.

The current study tests a novel prediction of the model about the role of uncertainty about personal attributes in evaluations of source reliability: Because traits that are central to one's self-concept tend to be associated with strong existing beliefs, feedback about these central traits has an especially strong influence on perceptions of source reliability. This prediction is tested in a behavioral experiment in which people evaluate the reliability of an "AI model" that generates predictions about personal attributes that vary in self-concept centrality.

Modeling Source Reliability

The present study extends previous work on modeling source reliability under a Bayesian framework (Pilgrim et al., 2024; Merdes et al., 2021). Past work has shown that a key driver of self-verification is the desire for accurate feedback (Swann et al., 2012; Swann, Stein-Seroussi, & Giesler, 1992). For attributes like personality traits there is no objective ground truth that might help establish the credibility of a source. A person can, however, use their subjective self-beliefs to evaluate source reliability. Feedback that validates existing self-beliefs should both reinforce those views and enhance the perceived reliability of the source. In contrast, disconfirmatory feedback that conflicts with one's self-views should raise suspicion about the source, especially when those self-views are held with high confidence. In this section I describe the application of the model to evaluations of either confirmatory (self-verifying) or disconfirmatory (non-self-verifying) feedback about attributes held with varying degrees of confidence.

The model assumes an individual maintains beliefs about a set of personal attributes, which in the current study represent various personality traits (e.g., "friendly", "rude", "self-centered"). For simplicity, each attribute is a binary property of either having the trait ($A = 1$) or not ($A = 0$), and the individual's degree of belief about whether they have the trait is represented by the prior $p(A)$. As described above, I assume that there are systematic differences in the strength of prior beliefs across attributes: Traits which have **high centrality** will tend to have stronger prior beliefs ($p(A)$ close to 0 or 1) compared to **low centrality** traits for which prior beliefs are more uncertain ($p(A)$ closer to .5).

The model separately maintains a belief about the reliability R of the feedback source, which is also represented as a binary property. When $R = 0$, the source is viewed as having no ability to provide accurate feedback about personal attributes and generates feedback at random. When $R = 1$, the source is viewed as being perfectly accurate in providing feedback that matches the truth about whether one has the attribute. Feedback is presented as a statement about whether the recipient has the attribute ($D = 1$, e.g., "You are friendly") or does not ($D = 0$, e.g., "You are NOT friendly"). The likelihood of observing each statement thus depends on both the underlying truth about the attribute (A) and whether the feedback source is reliable (R) (see Figure 1).

After a piece of feedback is observed, beliefs about both A and R are updated simultaneously in accordance with Bayes rule to determine the posterior distribution $P(A,R|D)$:

$$P(A,R|D) = \frac{P(D|A,R)P(A)P(R)}{P(D)} \tag{1}$$

The marginal posterior for source reliability $P(R|D) = \sum_A P(A,R|D)$ then serves as the prior when subsequent feedback is observed from the same source. This follows the independence assumption of the BIASR model from Pilgrim et al. (2024), in that the observer does not fully represent the joint belief distribution between source reliability and the entire set of attributes for which feedback has been observed (see Discussion).

Simulation

A simulation was performed to examine how beliefs about source reliability are influenced by the type of feedback received. There were two factors of interest. The first factor was whether the source generated feedback that confirmed or disconfirmed the beliefs held by the individual. Confirmatory ("self-verifying") feedback meant that the source generated a positive prediction ($D = 1$) when the individual believes they are more likely than not to have the target trait ($p(A) \geq .5$), while disconfirmatory feedback meant that the source generated a negative prediction ($D = 0$). For simplicity the simulation only included cases where $p(A) \geq .5$ since the effects of confirmatory and disconfirmatory feedback would be symmetric for beliefs in the opposite direction.

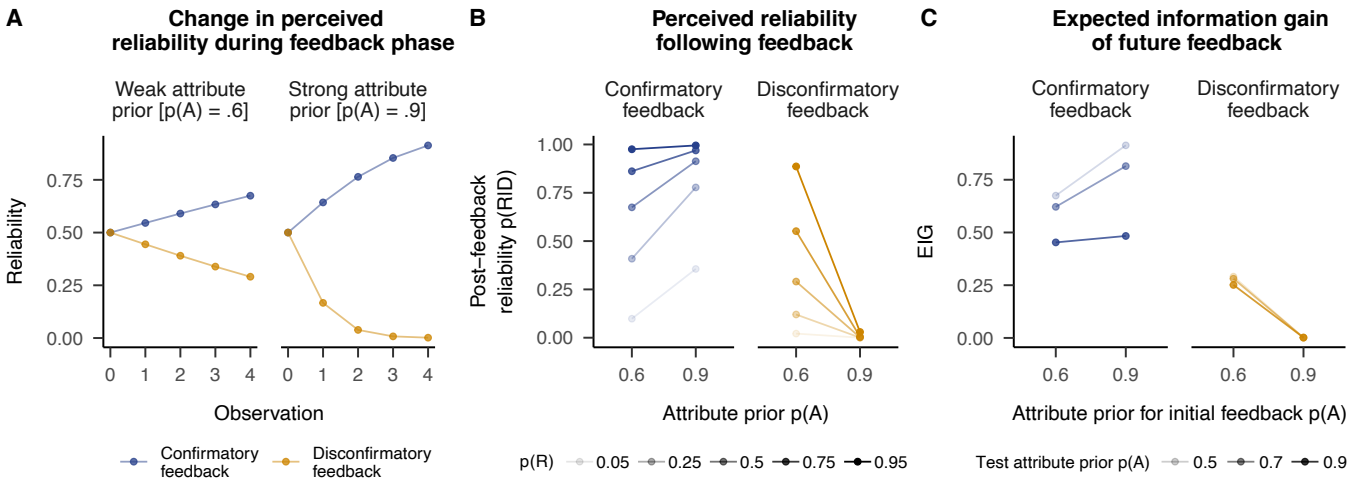


Figure 2: Simulation results. *A:* Example of change in belief about source reliability across four pieces of observed feedback, for an initial perceived reliability of $p(R) = .5$. Reliability increases with confirmatory feedback and decreases with disconfirmatory feedback, but these changes are smaller when prior beliefs about the feedback traits are weak (left) compared to when they are strong (right). *B:* Simulation results showing interaction between prior beliefs about attribute and feedback type on perceived reliability, for different levels of initial reliability $p(R)$. *C:* Expected information gain (EIG) of further feedback about attributes with varying degrees of prior belief ($p(A)$), after already receiving either confirmatory or disconfirmatory feedback about traits with either low ($p(A) = .6$) or high ($p(A) = .9$) prior belief. EIG is lowest after receiving disconfirmatory feedback due to low perceived reliability of the source, while EIG is highest after receiving confirmatory feedback about traits with high prior belief.

The second factor was the strength of prior beliefs $p(A)$ for attributes that were the basis for feedback. I compared prior beliefs that were either relatively uncertain ($p(A) = 0.6$, representing low-centrality traits) or highly confident ($p(A) = .9$, representing high-centrality traits). The expectation was that, in the absence of any ground truth against which to judge the feedback generated by the source, stronger prior beliefs would lead to larger changes in the perceived reliability of the source for both confirmatory and disconfirmatory feedback.

The simulation was conducted for different levels of initial perceived reliability, $p(R) \in \{.05, .25, .5, .75, .95\}$. Each run simulated the presentation of feedback for four separate attributes which all had the same prior belief, $p(A) \in \{0.6, 0.9\}$. The feedback for all four attributes was either confirmatory or disconfirmatory.

Figure 2A shows an example of how perceived reliability changes as each piece of feedback is observed, for an initial reliability of $p(R) = .5$. Overall, perceived reliability of the source increases following confirmatory feedback and decreases following disconfirmatory feedback. However, the strength of prior beliefs about the attributes moderates this effect: Changes in reliability are smaller when feedback is about more uncertain attributes ($p(A) = .6$) compared to when the feedback is about attributes for which the model is already highly confident ($p(A) = .9$). Figure 2B shows the final perceived reliability after observing feedback about four attributes across different levels of initial reliability $p(R)$, illustrating that this interaction between feedback type and the strength of prior beliefs is a general pattern regardless of what the individual initially believes about the feedback source.

These disparities in perceived reliability would be expected to influence whether a person is likely to seek out further feedback from the same source. Figure 2C displays the expected information gain (EIG; Nelson, 2005) from receiving an additional piece of feedback from sources with different perceived reliability after already observing feedback about other traits. EIG describes the reduction in uncertainty (change in entropy over the full posterior distribution) that is expected from receiving feedback from a source, weighted by the probability of each potential outcome. Results are shown for a new set of traits which vary in uncertainty, with $p(A)$ ranging from 0.5 (a highly uncertain attribute) to 0.9 (a highly certain attribute). After initial confirmatory feedback bolsters the perceived reliability of the source, the model predicts that further feedback about a new trait will be viewed as informative, especially for a trait with high uncertainty ($p(A) = .5$). Moreover, EIG is greatest after receiving confirmatory feedback about traits with greater self-certainty. In contrast, EIG is much lower when there is doubt about the source's reliability due to observing disconfirmatory feedback.

In sum, the simulation results suggest that because early self-verifying feedback about high centrality traits bolsters the perceived reliability of the source, an observer may then be more open to feedback about more uncertain traits from the same source. Early disconfirmatory feedback, however, leads to a rapid decrease in perceived reliability and would cause additional feedback from the same source to be seen as uninformative.

Experiment

A behavioral study was conducted to test the key model prediction that the perceived reliability of a feedback source depends on an interaction between feedback type (confirmatory vs. disconfirmatory) and the strength of prior beliefs about one's personal attributes. Participants completed an online task in which they first rated themselves on a set of traits that differed in their valence (positive vs. negative) and their centrality (low vs. high). Later in the task they received feedback from an "AI agent" that was described as having the ability to make predictions about their personality based on their responses to earlier questions. After viewing a set of predictions, participants made judgments about the AI system's reliability and their interest in viewing further feedback.

Participants

Participants were recruited via Prolific. $N = 169$ people completed the study (age $M = 37.8$ years, SD = 11.2; 44% female, 55% male, 1% no sex indicated). Participants received $4 for successful completion of the study. Eight participants were excluded from analysis because they completed one or more open-ended questions at a rate faster than 100 words per minute. Five additional participants were excluded because they failed more than 2 attention checks during the feedback phase. After exclusions data from $N = 156$ participants was retained for analyses, with between 37 and 44 participants in each of the four conditions (see below).

Materials and Procedure

Traits Elder et al. (2023) constructed a trait network based on people's judgments of the pairwise dependencies between traits. Sixteen traits (Table 1) were selected from that dataset which varied in perceived valence (positive or negative) and network centrality (low vs. high out-degree centrality).

Trait self-ratings Participants were first asked about each of the 16 traits in Table 1. They indicated the extent to which a trait described themselves on a scale from 0 (*Definitely does NOT describe me*) to 100 (*Definitely describes me*) using a slider. They then rated the extent to which it was personally important to be viewed by other people as having the trait on a 7-point scale (-3: *Very important to NOT be viewed in this way*; 0: *Unsure*; +3: *Very important to be viewed this way*).

Table 1: Traits organized by valence (positive/negative) and centrality (low/high).

	Positive traits	Negative traits
Low centrality	eager	cheerless
	poised	finicky
	prudent	insolent
	versatile	listless
High centrality	friendly	boring
	knowledgeable	reckless
	open-minded	rude
	respectful	self-centered

Finally, they rated the valence of the trait by indicating the extent to which it was a positive or negative way of describing a person (-3: *Very negative*; 0: *Unsure*; +3 *Very positive*).

AI predictions Participants were then instructed that the aim of the study was to evaluate an AI model's ability to predict their personal characteristics based on their responses to typical interview questions. They responded to three open-ended questions styled after standard job interviews (e.g., *"What is your greatest weakness? How have you taken steps to address it in the past?"*). Responses were required to be a minimum of 500 characters, related to the prompt, and completed at a rate of fewer than 100 words per minute.

After completing the open-ended questions, participants were told that their responses were provided to the AI model. They then proceeded to the feedback phase, during which they observed 4 predictions from the AI model about whether they had a certain attribute. Predictions for the four traits (two positive, two negative) were presented on separate screens in random order. A positive prediction was *"Yes, you are <trait name>."* A negative prediction was *"No, you are NOT <trait name>."* An attention check was included with each prediction where participants had to type in the trait name and select the matching prediction from a list of options.

Two factors were factorially manipulated:

- **Trait centrality**: In the **Low-centrality** condition, the 4 traits used in the feedback phase were selected from the low centrality traits in Table 1. Two positive and two negative traits were chosen if they had the lowest confidence (i.e., the participant's self-ratings were closest to 50). In the **High-centrality** condition, the 4 traits were selected from the pool of high centrality traits. Two positive and two negative traits were chosen which had the highest confidence (i.e., the participant's self-ratings were closest to 0 or 100).
- **Feedback type**: In the **Confirmatory** feedback condition, the AI prediction was always consistent with the self-rating given by the participant at the start of the study. Specifically, if the self-rating was greater than or equal to 50 the AI predicted the participant had the trait, otherwise it predicted they did not. In the **Disconfirmatory** condition, the AI predictions were reversed such that the prediction always differed from the direction of participants' self-ratings.

Perceived reliability. Participants rated the perceived reliability of the AI model both before and after the presentation of the model predictions. Participants rated perceived reliability at both timepoints using a slider from 0-100. They were instructed that the lower endpoint corresponded to "no ability" to predict their traits as if making predictions at random, while the rightmost endpoint represented "perfect accuracy" in predicting their traits.

Interest in further feedback. At the end of the task participants indicated their interest in viewing the AI model's

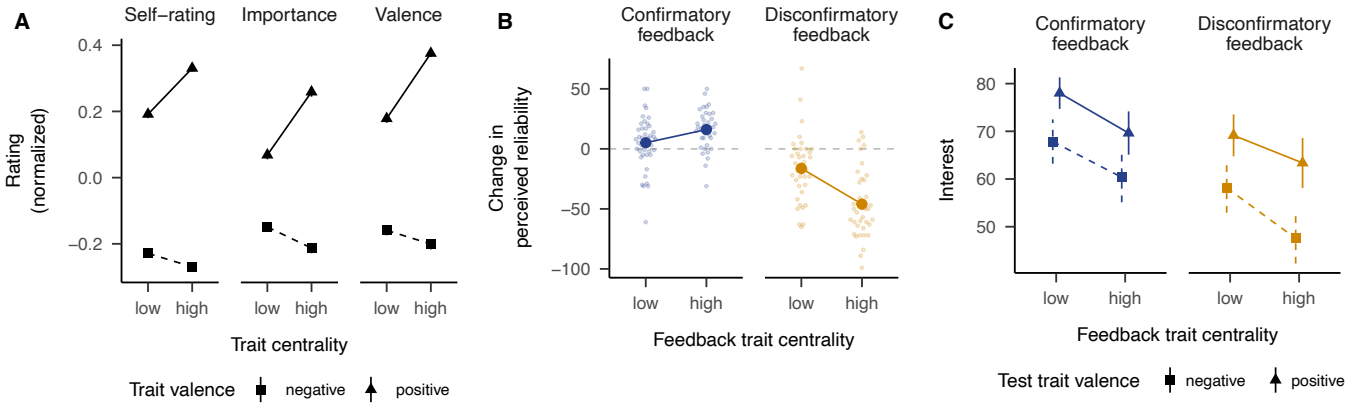


Figure 3: *A:* Participant ratings for all 16 traits separated by valence and centrality. *B:* Change in perceived reliability of the feedback source after receiving feedback. *C:* Interest in observing additional feedback about test traits.

predictions for a final set of 4 high-centrality traits that had not appeared during the feedback phase. Traits chosen for this phase were high centrality but lower confidence, and which participants may have been especially motivated to learn about since they tended to be viewed as important but also less certain.

Results

Trait ratings. For each of the 16 traits in Table 1 participants gave a self-rating, an importance rating, and valence rating (Figure 3A). Two-way mixed ANOVAs indicated significant main effects of valence and centrality as well as significant interactions for all three measures (all p's $< .001$). Post-hoc comparisons confirmed that responses to all questions were more extreme for high-centrality compared to low-centrality traits when the valence was either positive (all p's $< .001$) or negative (all p's $\leq .02$). This result suggests that participants had stronger self-beliefs for high-centrality attributes, viewed them as more connected to their identity, and had stronger positive or negative perceptions of their desirability. Although the present study did not include any steps to verify the centrality of these traits (e.g., by assessing trait dependencies as in Elder et al., 2023), these results corroborate the expected difference in the strength of prior beliefs between low- and high-centrality traits.

Perceived reliability. The main outcome of interest was how the perceived reliability of the AI model changed as a result of observing its predictions about personal attributes. A two-way ANOVA indicated significant main effects of feedback type ($F(1,152) = 118.29$, $p < .001$) and trait centrality ($F(1,152) = 6.02$, $p = .02$) as well as a significant interaction ($F(1,152) = 28.22$, $p < .001$). As seen in Figure 3B, perceived reliability was higher overall following confirmatory feedback compared to disconfirmatory feedback. Trait centrality had different effects depending on the type of feedback. For confirmatory feedback, perceived reliability increased more when that feedback was about high-centrality traits ($t(152) = 2.06$, $p = .04$). For disconfirmatory feedback, perceived reliability decreased more when feedback was about high-centrality traits ($t(152) = 5.40$, $p < .001$). This aligns with the predicted interaction between trait centrality and feedback type from the model (see Figure 2B).

Interest in further feedback. In the final stage, participants indicated their interest in receiving additional feedback about four high-centrality traits (two positive and two negative) that did not appear in the feedback phase. Responses were averaged for traits of the same valence (Figure 3C). A mixed effects model was run with fixed effects for trait valence, trait centrality, feedback type, and the interaction between trait centrality and feedback type. Random intercepts were included for participants and traits. There was a significant effect of feedback type ($t(151.54) = 2.23$, $p = .03$) such that interest was higher following confirmatory than disconfirmatory feedback, aligning with the expected effect from the model simulations (Figure 2C). There was also a significant effect of valence ($t(5.52) = 4.86$, $p = .004$) such that interest was higher for positive than negative traits. No other effects were significant. As shown in Figure 3C, interest in further feedback was lowest overall for negative traits after participants had already observed disconfirmatory feedback about high-centrality traits.

Discussion

The way that people seek out and evaluate social feedback has important implications for their self-image. Research on self-verification has shown that people often prefer feedback that aligns with existing self-views, even when other sources of disconfirming feedback are available (Swann et al., 2012). The present study adopted a formal approach for modeling beliefs about source reliability (Pilgrim et al., 2024; Merdes et al., 2021) in order to better understand how the preference for self-verification is rationally motivated by epistemic concerns about the credibility of feedback given varying degrees of uncertainty about one's own traits.

The results of the behavioral experiment confirmed a key prediction of the model: Confirmatory (self-verifying) feedback and disconfirmatory feedback have opposing effects on the perceived reliability of the source, and this effect is magnified when feedback is offered on high centrality traits that tend to be associated with strong prior beliefs and viewed as important to one's self-concept. This helps explain why self-verification is more likely when source reliability is in doubt compared to when observers expect the source to be an "expert" based on other cues (Szumowska et al., 2022). Note that while the current report focuses on group-level differences using fixed trait categories, future analyses will explore whether individual-level variation in perceived reliability is predicted by participants' subjective self-views about the traits for which feedback was given.

The basic model described here can be extended in future work to better understand the evolution of trust in a source across several opportunities to receive their feedback. The current study focused on simple comparisons between sources that provided one type of feedback (either entirely confirmatory or disconfirmatory) for similar sets of traits (either entirely low- or high-centrality). Real-world interactions with feedback sources likely involve more complex combinations of both factors. For instance, a therapist may at various points provide feedback which either affirms or challenges an individual's existing self-views. They might also focus on low-centrality beliefs that are more malleable before addressing more deep-seated, maladaptive self-beliefs. In that context, self-verifying feedback—even if it affirms some negative self-beliefs—may be critical to establishing trust and sustaining engagement in therapy (Swann, 1997).

An important consideration for future work is how beliefs about dependencies between personal traits factor into judgments of source reliability. In the present study, self-beliefs were treated as independent hypotheses about whether one has a particular trait. However, other work shows that self-knowledge is highly organized, with people viewing some traits as being more or less dependent on others. From a network perspective, high-centrality traits are those upon which many other traits depend and which are key to maintaining the stability of one's self-concept (Elder et al., 2023). One implication of this is that feedback that violates these dependencies may be especially damaging to a source's reliability (e.g., being told that one is "friendly" but also "NOT approachable"). These relationships are not captured by the current model, which follows the independence assumption from Pilgrim et al. (2024) that the observer does not represent the full joint belief distribution between source reliability and all of the attributes about which feedback was provided. Pilgrim et al. (2024) motivate this assumption as a necessary approximation to rational inference given that tracking the full distribution would typically be computationally infeasible. Moreover, they show that this approximation leads to order effects such that early pieces of evidence can have outsize influence on perceived reliability and lead to confirmatory biases in the evaluation and selection of later evidence. Future work will examine whether this independence assumption is appropriate for feedback about interrelated personal attributes and whether similar order effects occur for self-verifying feedback.

Finally, the present approach lays the groundwork for a formal understanding of how people make decisions between sources of self-relevant feedback. The model simulation showed that, once some initial feedback is observed, the perceived reliability of the source has a strong influence on whether additional feedback is likely to be seen as informative (Figure 2B). This prediction was consistent with the behavioral finding that participants in the Disconfirmatory condition were less interested in additional feedback compared to the Confirmatory condition. However, the results also indicated a strong effect of the valence of test traits, with participants showing greater interest in additional feedback about positive traits than negative traits, despite the fact that both kinds of traits were rated with similar degrees of self-certainty and importance. This preference for feedback about positive traits could reflect a desire for "self-enhancing" feedback that portrays oneself in a positive light (Alicke & Sedikides, 2009) or the need to manage one's emotions or self-esteem (Glass, Levens, & Markant, under review).

Higher interest in confirmatory sources may have also been related to expectations about whether further feedback was likely to be favorable. Although feedback was presented about both positive and negative traits, it is important to note that confirmatory feedback tended to be more desirable compared to disconfirmatory feedback, as participants largely endorsed positive self-beliefs that were affirmed by the AI system (whether that entailed having positive traits or *not* having negative traits). Source reliability and the favorability of feedback could be disentangled in future work by targeting negative self-views. Prior work suggests that people with low self-esteem or depression often favor self-verifying feedback that reinforces negative self-views (Swann, Wenzlaff, et al., 1992), a preference which has been linked to doubts about the credibility of positive, disconfirming feedback (Swann, Stein-Seroussi, & Giesler, 1992). An important problem for future work is to understand how early feedback should be designed in order to foster a willingness to engage with favorable feedback that could potentially alter negative self-views. While the present work focused on the simple relationship between perceived reliability and information value, a natural extension would be to consider other dimensions of utility (including affective consequences) that influence how people perceive and pursue feedback from different sources.

In real-world settings, a range of personal and situational factors may impact whether people view feedback as personally relevant or useful in the moment (Szumowska, Szwed, Wójcik, & Kruglanski, 2023). Formal modeling of these decisions may lead to a better understanding of the adaptive roles of self-verification in cultivating accurate self-assessment, positive self-image, and interpersonal trust.

Acknowledgments

The author is grateful to the anonymous reviewers for valuable feedback which was used to improve the paper.

References

Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology*, *20*(1), 1–48. doi: 10.1080/10463280802613866

Anseel, F., Lievens, F., & Levy, P. E. (2007). A self-motives perspective on feedback-seeking behavior: Linking organizational behavior and social psychology research. *International Journal of Management Reviews*, *9*(3), 211–236. doi: 10.1111/j.1468-2370.2007.00210.x

Elder, J., Cheung, B., Davis, T., & Hughes, B. (2023). Mapping the self: A network approach for understanding psychological and neural representations of self-concept structure. *Journal of Personality and Social Psychology: Attitudes and Social Cognition*, *124*(2), 237–263. doi: 10.1037/pspa0000315

Glass, S., Levens, S., & Markant, D. (under review). Adaptive emotion regulation and the preference for self-verifying feedback.

Hahn, U., Merdes, C., & Von Sydow, M. (2018). How good is your evidence and how would you know? *Topics in Cognitive Science*, *10*(4), 660–678. doi: 10.1111/tops.12374

Jern, A., Chang, K. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, *121*(2), 206–224. doi: 10.1037/a0035941

Merdes, C., Von Sydow, M., & Hahn, U. (2021). Formal models of source reliability. *Synthese*, *198*(S23), 5773–5801. doi: 10.1007/s11229-020-02595-2

Nelson, J. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *114*(3), 677. doi: 10.1037/0033-295X.112.4.979

Nickerson, RS. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220. doi: 10.1037/1089-2680.2.2.175

Pilgrim, C., Sanborn, A., Malthouse, E., & Hills, T. T. (2024). Confirmation bias emerges from an approximation to Bayesian reasoning. *Cognition*, *245*, 105693. doi: 10.1016/j.cognition.2023.105693

Swann, W. (1997). The trouble with change: Self-verification and allegiance to the self. *Psychological Science*, *8*(3), 177–180. doi: 10.1111/j.1467-9280.1997.tb00407.x

Swann, W., Rentfrow, P., & Sellers, J. (2012). Self-verification: The search for coherence. In *Handbook of Self and Identity, 2nd ed.* (p. 405 - 424). The Guilford Press.

Swann, W., Stein-Seroussi, A., & Giesler, R. (1992). Why people self-verify. *Journal of Personality and Social Psychology*, *62*(3), 392. doi: 10.1037/0022-3514.62.3.392

Swann, W., Wenzlaff, R. M., Krull, D. S., & Pelham, B. W. (1992). Allure of negative feedback: Self-verification strivings among depressed persons. *Journal of Abnormal Psychology*, *101*(2), 293. doi: https://doi.org/10.1037/0021-843x.101.2.293

Szumowska, E., Szwed, P., Wójcik, N., & Kruglanski, A. W. (2023). The interplay of positivity and self-verification strivings: Feedback preference under increased desire for self-enhancement. *Learning and Instruction*, *83*, 101715. doi: 10.1016/j.learninstruc.2022.101715

Szumowska, E., Wójcik, N., Szwed, P., & Kruglanski, A. W. (2022). Says who? Credibility effects in self-verification strivings. *Psychological Science*, *33*(5), 699-715. doi: https://doi.org/10.1177/09567976211049439