Assessment of actively open-minded thinking

The explicit concept of reflective thought as something to be encouraged is fairly recent in the development of human cultures. Surely it emerged more than once, but in Western culture it seems to have originated with the Greeks, a few hundred years before Christ. In the writings of philosophers, it has been discussed more or less continuously since them. One high point was the influential *Port Royal Logic* of 1662 (Arnauld, 1964).

John Stuart Mill was perhaps the clearest 19th century advocate of reflective thought. In *On liberty* (1859, ch. 2), he writes (as part of a longer argument): "The whole strength and value, then, of human judgment, depending on the one property, that it can be set right when it is wrong, reliance can be placed on it only when the means of setting it right are kept constantly at hand. In the case of any person whose judgment is really deserving of confidence, how has it become so? Because he has kept his mind open to criticism of his opinions and conduct. Because it has been his practice to listen to all that could be said against him; to profit by as much of it as was just, and expound to himself, and upon occasion to others, the fallacy of what was fallacious." Other philosophers have also made much of this theme, especially John Dewey (1901/2012) and John Rawls, who discussed "reflective equilibrium" as a method of inquiry in philosophy itself.

Several researchers have developed this idea into a concept of actively open-minded thinking (AOT). The introductory section of this proposal reviews the nature of AOT, and distinguishes AOT from reflection in general, or "system 2" reasoning. I then review some attempts to measure AOT and their problems, which will lead to the proposed research. The research proposed will develop different ways of measuring AOT, including: questionnaires about the nature of good thinking, self-report questionnaires, simulated decision making, and grading and analysis other people's thinking. These methods will be validated by looking at tests of generality across domains, tests of transfer when training is provided using some of the content areas, the correlations of different methods with each other, and their correlations with other criteria, including criteria for belief, moral judgment, and overconfidence in probability judgment. In the end, we hope to have more than one useful scale, a prototype for training AOT, and a deeper understanding of AOT itself as a trait and a property of judgment and decision making.

Actively open-minded thinking (AOT)

AOT is the disposition to be fair towards different conclusions even if they go against one's initially favored or pet conclusion. In an early demonstration, Perkins, Bushey and Faraday (1986) asked students to write down their thoughts on issues that were "genuinely vexed and timely" and that could be discussed on the basis of knowledge that most people have, e.g., "Would providing more money for public schools significantly improve the quality of teaching and learning?" Most students gave more arguments on their favored side, "myside" thoughts, than on the other side. When the students were asked to try harder to think of arguments on each side, they thought of very few additional myside arguments but many additional otherside arguments. Left to their own devices, then, the students looked primarily for reasons to support their initial opinion, but out of biased search rather than lack of ability or knowledge.

Baron (1985, 1988–2008), proposed a general framework for discussing thinking in terms of search for "possibilities, evidence and goals" and making inferences from these. He also outlined a general theory of where thinking often goes wrong, specifically, in failing to search for possibilities and goals other than those "on screen" at the moment, failing to look hard enough for evidence against favored possibilities, and under-weighing evidence against favored possibilities when it is available. Baron called this set of deficiencies "myside bias." The general set of dispositions that would reduce these biases was called "actively open-minded thinking" (AOT). AOT is not merely being open to reasons why a favored possibility might be wrong but also actively looking for them, in the spirit of J. S. Mill's ideal. Of course, this whole approach drew heavily on earlier work, particularly that of Irving Janis and collaborators (e.g., Herek, Janis & Huth, 1987), who developed a similar framework for analysis of decision making in particular.

Other approaches to reflective thinking

AOT is not the same as general reflectiveness in thinking. Two other traditions deal with reflectiveness. One is the dual-system theory proposed by many researchers (e.g., Evans, 2008; Evans & Stanovich, 2013; Kahneman, 2011; Sloman, 1996; Wason & Evans, 1975), in which an intuitive system produces conclusions, which may be examined more thoroughly by a reflective system, which (in most accounts) begins its operation after the intuitive system has generated an initial conclusion. Insofar as this theory fits the data from actual thinking, system two may or may not be critical of initial conclusions. It is also used to bolster initial conclusions even when they are completely incorrect (e.g., Wason & Evans, 1975). Haidt (2001) has argued that such bolstering is also common in thinking about moral questions, where an initial emotional response is justified by subsequent thought, rather than questioned. Thus, dual-system theory is technically neutral with respect to the most important feature of AOT, the direction of thinking with respect to conclusions that are in favor at the moment. Empirically, system-two reflection does serve to correct errors in many situations (e.g., Johnson-Laird & Bara, 1984), so we would expect the use of system two to be correlated with AOT, but it is not the same, and measures of the use of system two would not serve as sufficient measures of AOT.

A second relevant tradition is the study of reflection-impulsivity (R/I) as a dimension of individual differences in problem-solving tasks (Kagan, Rosman, Day, Albert, & Phillips, 1964; Kagan, Pearson & Welch, 1966; reviewed by Messer, 1976, and by Baron, Badgio & Gaskins., 1986,), such as a complex visual-matching task. "Reflective" children are those who choose to be careful at the expense of speed. Impulsive children do things quickly but make many errors. Speed and accuracy are often negatively correlated. The "impulsives" are the children whose answers are fast and inaccurate; the "reflectives" are the children whose answers are the reverse. The test is useful for prediction. Reflectives tend to be older, to score higher on IQ tests (even when the tests are timed), and to be less prone to disruptive behavior in the classroom.

R/I can be measured in any task in which accuracy and response-time (RT) can be measured. It makes sense, though, to limit the measure to novel tasks that require thinking and do not have a natural time limit. In such tasks, it is often possible to find a positive correlation between RT and accuracy. In many studies, it turns out that log(RT) is consistent across tasks and is sometimes just as useful a predictor as accuracy. For example, log(RT) of items from the Raven's Progressive

Matrices, as well as accuracy on these items, predicts forecasting accuracy (Mellers et al., 2015). Baron, Scott, Fincher and Metz (2015) also reported both cross-task consistency of RT and correlations with measures of utilitarian moral judgment, (dis)belief in divine-command theory (described later), and a measure of AOT. Baron et al. (1986) report a *positive* correlation between latency and a measure of IQ, as well as evidence of consistency across tasks.

R/I should also be correlated with the use of system two, but it need not be. Alternatively, it could be related to a person's initial approach to a task. For example, in the famous "bat and ball problem", a person may either guess the answer and then check it (or not) or may apply simple algebra from the outset.

Baron et al. (2015) argue that at least some of the predictive power of R/I is that it is correlated with AOT. The two concepts are not the same, but they overlap. R/I is concerned with thoroughness in search, but it does not assess the direction of search. Long RTs could result from searching for reasons favoring the initially favored possibility. R/I is a convenient measure; however, AOT offers a fuller picture of good thinking; depth can be inadequate without breadth of direction.

Measurement of AOT

Baron (1991, 1995) argued that beliefs affect what people do, and supported this with correlations between judgments about what good thinking is and the subjects' own thinking. Stanovich and West (1997, 1998) found additional supporting evidence. They constructed a Thinking Dispositions Questionnaire, with items from many other scales that captured the general idea of AOT (such as the Need for Cognition scale reviewed by Cacioppo, Petty, Feinstein & Jarvis, 1996), plus a new part, the Flexible Thinking scale, which they designed themselves specifically to measure AOT as defined by Baron. In subsequent papers over the years, this scale was revised. A later version used by Toplak, West and Stanovich (2014) contained 18 items.

Several papers (reviewed by Toplak et al., 2014, and Stanovich, 2016, Table 1), found correlations between this scale and other tests. One, the Argument Evaluation Test, looked directly for myside bias in the evaluation of arguments on both sides of an issue. Many others measured biases described in the literature on judgment and decision making, including (but not limited to): Baserate Neglect, Conjunction Fallacy, Framing Effects, Anchoring Effect, Sample Size Awareness, Regression to the Mean, Temporal Discounting, Gambler's Fallacy, Probability Matching, Overconfidence Effect, Outcome Bias, Ratio Bias, Ignoring P(D/ H), Sunk Cost Effect, Risk/Benefit Confounding, Omission Bias, Expected Value Maximization, Hindsight Bias, Certainty Effect, Willingness to pay/Willingness to accept, and Proportion Dominance Effect. Stanovich and his colleagues have put together a battery of these tests to form a Comprehensive Assessment of Rational Thinking, which is designed to measure important cognitive dispositions that are omitted from most IQ tests.

Baron selected items from the original questionnaire and added others to make a different short form of the AOT scale, designed to assess beliefs about the nature of good thinking in particular. Example items are: "People should take into consideration evidence that goes against their beliefs."; and "Changing your mind is a sign of weakness." This short form has had considerable success in predicting the results of other tasks (Haran, Ritov & Mellers, 2013; Mellers et al., 2015; and Baron et al., 2015, using a slightly extended version).[1] Yet, this is only a measure of beliefs about good

[1]The current version is at http://sjdm.org/dmidi/Actively_Open-Minded_Thinking_Beliefs.html.

thinking. The fact that it correlates with task performance suggests that efforts to change these beliefs, when they oppose AOT, could result in improved performance.

Metz (2017) has developed a related scale for adolescents, the *Adolescent Actively Open-Minded Thinking Scale (AAOT)*. The novel scale was piloted one-on-one and in small groups with students ages 12–13 to ensure adolescents would interpret the items as intended, and adjusted accordingly. The AAOT exhibited a Cronbach's alpha of .72 (N=1356) and a six-month test-retest validity of .59 (N=1306). Consistent with the construct, the AAOT scale predicted an empathy scale ($r=.41$, $p<.001$, N=1185) and a perspective-taking scale ($r=.42$, $p<.001$, N=1302). It also predicted a number of non-self-report, more behavioral measures of actively open-minded thinking, based on interviews.

Earlier studies are consistent with the conclusion that beliefs about the nature of good thinking correlate with more direct measures of thinking itself. Perkins, Bushey and Faraday (1986) and Baron (1995) both found correlations between measures of beliefs about thinking and myside bias as measured by asking people to give arguments about some potentially controversial issue.

The occasional low scores on this scale are somewhat surprising, since the items almost seem to be measuring social desirability. Who could be against open-mindedness? One possible answer, suggested in Baron (2008) and supported in Baron et al. (2015) is that some people grow up in cultures that oppose questioning, lest children come to question doctrines dictated by authority. Baron et al. (2015) find large negative correlations between AOT and a measure of belief in "divine command theory" (Piazza & Landy, 2013), the idea that people do not have the capacity to engage in moral reasoning or to understand it, therefore, we must accept the word of God or some other authority without question and without understanding. It would appear that some cultural institutions, in order to prevent questioning of their authority, tend to inculcate the belief that thinking, curiosity, and questioning are more generally undesirable.

The theory of AOT

Here I present a sketch of the theory of AOT, as described at greater length in Baron (1985, 1988–2008). The theory is meant to apply to all types of thinking, but we may divide types into three categories: thinking about beliefs, goals, and options. Thinking about options is "decision making", but it can include more complex episodes such as artistic creation or "dynamic decision making".

Thinking can be described in terms of search and inference. The objects searched for consist of possibilities, evidence (in the form of beliefs) and goals (criteria, values, desires). In decision making, the possibilities are options, and inference is the evaluation of options in the light of evidence and goals. We can also think about beliefs themselves, by evaluating them in terms of other beliefs. (Goals may play a role too, but this is a distortion.) And, similarly, we can evaluate goals in terms of other goals. Most of the thinking studied in the relevant psychology literature is about decisions and beliefs. Goals have received little attention in psychology (but more from philosophy).

In general, episodes of thinking may be characterized in terms of a few very general parameters. Search may be characterized by its objects (evidence, possibilities, or goals), its duration or extent (time, or number of objects sought), and, importantly, its direction. Direction is defined in terms of whether search is directed at whatever favors currently strong possibilities (usually just one) or opposes them. Inference may also be characterized by direction. For example, evidence may be weighed more heavily as a function of whether it favors or opposes strong possibilities.

A few concepts in the literature on individual differences are concerned with these general parameters. In particular, R/I (described above) is concerned with the extent of search, but not with its direction. Moreover "maximizing tendency" (Cheek & Schwartz, 2016) may be defined in terms of extent of search, and should thus be correlated with R/I. Maximizers, as distinct from satisficers, are those who engage in extensive search for options, when making decisions, or who have high standards for ending the thought process and choosing an option. These two criteria are relevant in different situations. Extensive search for options is relevant when many options are available (e.g., when searching for a movie to watch on Netflix). When only a few options are being considered, extensive search for evidence is also possible (e.g., reading IMBD reviews after selecting a few candidate movies). High standards are relevant when the decision of taking an option is possible, vs. waiting for a better option to come along, or not choosing any option in the relevant set (e.g., deciding not to watch a movie, because one meeting the high standard [including not being seen before] cannot be found). Low standards are usually incompatible with extensive search, because an option meeting these standards can be found quickly. In principle, then the two important components of maximization should be correlated across people with reflection-impulsivity, and hence with AOT.

The main conceptual contribution of AOT is its concern with direction, as well as extent. This is because AOT is intended as the antidote to myside bias. Like R/I, it is an Aristotelian virtue, because there is a happy medium, an optimum; it is possible to search too much, or to be too self-critical. In the latter case, people would be tied in knots over each decision, and the extra time spent deciding would not be compensated by the increased chance of a better outcome. But, in the case of AOT, departures from the optimum are more common on the side of myside bias. Thus, we call it a virtue when people resist myside bias by looking for reasons why their pet belief or favored option might be wrong. It is unclear whether reflectiveness (as part of R/I) or maximizing are virtues in the same way. In some cases people clearly think too much about issues that do not warrant such attention, or they feel compelled to find an answer when it is more reasonable to leave some question unanswered for the time being. Yet, it is also true that the search for arguments against a favored option does take some time and effort, so there ought to be (and apparently is) some correlation between AOT and R/I (Baron et al., 2015). Importantly, both dimensions vary considerably across people.

In the case of belief in particular, AOT ought to be reflected in appropriate confidence. If you have not looked for reasons why your favored belief might be incorrect, you should not have so much confidence that it is correct. A person high in the trait of AOT can say "I don't know the answer. I haven't looked into that question very thoroughly." One of the classic findings showing myside bias is that of Koriat, Lichtenstein and Fischhoff (1980), showing that instructions to think of reasons supporting a favored answer did not have much effect on confidence, presumably because subjects had already done that, but instructions to think of reasons on the other side reduced extreme confidence in incorrect answers. Gürçay-Morris (2016) also found that overconfidence could be reduced by instruction in AOT, and was correlated negatively with some measures of individual differences in AOT.

AOT is relevant both to the questioning of long-held beliefs or plans and to thinking about new problems arising for the first time. When conclusions are required, AOT requires sufficiently extensive search, within the constraints at hand, as well as fairness in its direction. Fairness in direction is possible even when time is limited. In sum, AOT manifests itself across several distinctions:

- amount of search vs. direction of search;
- direction of search vs. direction of inference;
- search for possibilities vs. evidence;
- conclusions required vs. not required (but confidence is assessed);
- solving new problems vs. questioning of existing beliefs and plans.

AOT in all of these cases is tied together by its role in insuring that thinking yields beliefs that are most warranted, confidence that is justified, and decisions that are as close to optimal as time permits. But we could well expect them to be imperfectly correlated with each other across these distinctions, as measures of individual differences. Thus, we should not expect any measure of AOT to be uni-dimensional.

Understanding AOT

Baron (1993) argued that AOT is important for two different purposes. One is that it is useful in achieving goals, and in reflectively formulating what those goals should be. The other, which is becoming more important as the world becomes more complex, is that it provides a criterion of which sources of advice to trust. This was also the emphasis in Mill's argument presented at the beginning of this proposal. Today, we cannot reasonably make our own decisions about health care without the advice of a medical doctor (putting aside the control that doctors have over access to medical interventions). People often are wise to consult lawyers and other professionals with specific knowledge; those of us who have tried household repairs have often learned the hard way that plumbers and electricians sometimes know more than we do. As citizens, we are influenced by sources of news, by politicians, by scientists, and by advocates of various views, including people we know personally. How do we know whether some influence on us is (in Mill's terms) a "person whose judgment is really deserving of confidence"?

Mill's answer, and mine (1993), is that such a person has reached conclusions through AOT. In some cases, the AOT is done by several people. For example, some scientists are defensive about their ideas, but science as an institution engages in AOT because individual scientists get credit for poking holes in the conclusions of other scientists. Thus, understanding the value of science as an institution requires an understanding of the nature of AOT. It is what distinguishes true science from pseudo-science. Science is not the only institution that works this way.

What does it mean to understand AOT? Following the ideas of Wertheimer (1959) and Perkins (1986), I suggest that the crucial realizations involve knowing the purposes of the various elements of AOT. For example, the reason we look for other possibilities is to make sure that the current favorite is really the best, or to look for ways to modify it to make it better, by taking pieces of other possibilities. The reason we look for counter-evidence is, again, to prevent error and to suggest ways to modify a possibility. More generally, the reason for all of these elements is to increase our justified confidence in whatever possibility we choose in the end. This last point is a little subtle, but important. It could be tested, for example, by asking, "How confident should we be in this conclusion if we do not [do X]?", where X is some type of thinking, and "If we [do X] and reach the same conclusion, what should happen to our confidence that this conclusion is the best choice?" Note that simply asking questions like this could induce reflection on the answer and thus serve as a

component of training in AOT. In any experiments other than those about training, they will be the last questions asked.

Importantly, it should be possible to design questions in this format that would not be trivially easy to answer, yet, at the same time, valid indicators of what needs to be understood, hence useful in a test. Such a test of understanding would have the advantage of having correct answers, hence being difficult to fake.

Problems with the AOT scale

Despite occasional low scores, most scores on AOT scales are high, especially the AAOT given to middle-school children. In a sense, these tests are too easy. It is easy to say that it is good to be open-minded, even actively open-minded. This is a socially endorsed virtue. (Stanovich & West, 1997, found no correlation with a social desirability scale, but it is possible that subjects in these experiments also figured out what the experimenters were looking for, so that they put themselves in the position of the experimenters, rather than society in general.)

Yet all forms of this scale leave out understanding, they do not allocate many questions to search as distinct from inference, they contain questions that refer more to side-effects of AOT and stereotypes of biased thinkers than to the mechanisms at issue, and they contain items implying that more is always better, in contrast to the concept of an Aristotelian virtus. For example (from different versions of the AOT scale, including mine):

- "A person should always think about new possibilities." The term "always" would cause hesitation in someone who understood the nature of AOT as a virtue, or even for someone who reflected carefully on the meaning of the statement.
- "People should always think about evidence that goes against their beliefs." Again, "always" is too strong. If you have already thought about it many times, it is wasteful to keep returning to the same issue.
- "Feelings are the best guide in making decisions." This is part of the stereotype, but unrelated to the definition. The term "feelings" is somewhat unclear; feelings could include justified feelings of coherence or rightness.
- "I already know mostly everything I need to know." This is unrelated to the definition. It has more to do with curiosity, except that it refers to "need".
- "People make bad choices when they listen to lots of different opinions." This could be true if they listen to fools.
- "People who criticize me often don't know what they are talking about." This depends heavily on who you are, and who the critics are.
- "Nobody can change my mind if I know I am right." Taken literally, "know" implies that the belief in question is true.
- "Most people just don't know what's good for them." This has nothing to do with the definition of AOT.
- "Wise people make fast decisions." If wisdom is the same as expertise, this is often true.
- "Intuition is the best guide in making decisions." This has little to do with the definition, and it could be true for the final step, in which all the evidence is put together. "Best" compared to what?

- "It really makes me angry when someone can't say they are wrong." This is reverse scored. I had to check. It isn't obvious. If you value AOT, then you might well be annoyed at people who don't do it. You may think that AOT is a *moral* virtue as well as an intellectual one.
- "I think people are either with me or against me." This seems to be based on a stereotype of paranoia.
- "When we are faced with a new question, the first answer that occurs to us is usually best." This could be true, empirically.

Of course, most of these items perform well in tests of both reliability and validity. But when a scale like this includes many questionable items, it is difficult to tell what it measures, and it may indeed be systematically sensitive to other dimensions than AOT.

Training AOT

Although the problem of measuring AOT efficiently is not fully solved, it is clear that AOT can be taught. To some extent, culture and education may do this. Indeed, to some extent it could be argued that most people would learn how to think well on their own, but for the barriers set up by cultures that survive because they discourage questioning. Curiosity, for example, seems to be a common feature of childhood, yet parents and teachers often discourage it by telling children that certain questions should not be asked.

Experiments in training AOT have been going on since the remarkable studies of Otto Selz (1935; summarized in Baron, 2008). More recently, Perkins et al. (1986) taught high school students to think in an actively open-minded way through a sixteen-session course that emphasized searching thoroughly for arguments on both sides of an issue. Controversial issues were discussed in class, and students were explicitly encouraged to generate and evaluate (for truth and relevance) arguments on both sides, especially the other side.

Before and after the course, students were tested by asking them to list arguments relevant to controversial issues (as described above). The course nearly doubled the number of *otherside* arguments. These gains were not simply the result of greater thoroughness in general: The course did not increase the number or quality of myside arguments. The effect was truly a matter of a change in direction. Perkins and his colleagues also examined the effects of other courses that involve thinking in some way: a first-year law-school class, a high school debate class, a first-year college class that taught "critical thinking," and a graduate course on thinking. None of these courses affected otherside arguments significantly, although the law-school class and the critical-thinking class increased the number and quality of myside arguments. In sum, it seems that we can successfully teach actively open-minded thinking by encouraging it directly, but not by simply requiring students to think.

Gaskins and Baron conducted an 8-month training study in her school for reading-disabled children, the Benchmark School (Baron, et al., 1986). The study was more designed to teach general reflectiveness, in the sense of R/I, than AOT, but it is worthy of note because the training affected R/I measures in various experimental tasks, as well as affecting teacher ratings, including ratings given by teachers of children who had to other schools.

More recently still, we have developed a training procedure that can be done on the World Wide Web[2], and a new behavioral measure to test its effects (Gürçay, 2016).

In the first part of the training subjects were informed about the nature of the training, and were presented with two questions: a percentage question where subjects were asked to make a point estimate by assigning a value between 0 and 100, and a multiple choice question where they had to choose one option out of three and to make probability judgments for each option's correctness. For both questions subjects were asked to explain how they came up with their answers.

In the second part of the training subjects read about what thinking is, and a discussion of important concepts such as "possibilities," "evidence," "goals," and "conclusion." They also read short paragraphs about what actively open-minded thinking is, why AOT is good thinking, and how it is useful. They were tested on understanding of these concepts and given feedback. Many answers were incorrect, indicating that our test of understanding describe above will not be too easy.

In the third part of the training subjects learned about myside bias and how this bias operates. Then they read the responses of two hypothetical respondents who modeled either myside bias or AOT for the two questions subjects answered at the beginning of the training. Subjects were explicitly told why and how each respondent was displaying myside bias or actively open-minded thinking. After looking at these modeled responses, subjects got to see their own responses to these two questions and the correct answers. They were also asked to self-evaluate their responses in terms of how much myside bias or principles of AOT they displayed. After completing this exercise, subjects took a three-item review test to evaluate how well they understood the concepts of AOT. They were then given feedback.

In the fourth and final part of the training, subjects were given six problems to think about, of which they had to answer 5 (two out of three policy questions and all three probability questions), listing relevant arguments. They then classified their arguments as "for" or "against" their favored solution or neither. After the classification task, they could change what their favored solution. Subjects then received feedback and evaluated their responses. At the end of the training subjects took a 7-item survey asking about their experience of learning about actively open-minded thinking.

To test the effect of the training, we gave subjects 20 difficult multiple-choice items, such as which of three cities is the largest in population. We asked subjects to list reasons, and probabilities, and to classify their own reasons as being for, or against, each of the three options. Our measure of AOT was the number of reasons against the option they chose (including those that were for some other option). Scores on this measure were strongly increased as a result of the training, between a pre-test and post-test. Importantly, training also reduced one measure of over-confidence as expressed in these judgments, although this effect was not as large as we hoped.

In sum, it seems possible to devise an effective web-based training module, which includes definitions of terms and tests of understanding along the way. A longer version with several lessons would surely create more lasting change, but it is possible that even the current version had a lasting effect on beliefs about what good thinking should involve.

[2]http://finzi.psych.upenn.edu/~baron/ex/bg/args/args10TrainingQualtrics.pdf.

Specific methods and proposed research

The proposed research has the following parts. First, design new scales concerning beliefs about good thinking and test the generality of these scales. In all cases, new items will be pilot tested and revised, both by asking paid subjects to comment on them and also by asking colleagues to read and comment on them.

As a possible alternative scale, we will ask subjects to assign "grades" to brief reports of someone else's thinking, and to indicate reasons for the grades (as done in a simple form by Baron, 1995).

Second, we will examine correlations of these scales with four types of criterion variables, as a test of validity:

1. Three-choice questions with correct answers, of the sort used by Gürçay (2016), with probability judgments. These will allow an assessment of overconfidence.
2. Simulations of thinking that track subjects through on-line thinking tasks, such as consumer purchases.
3. Scales of utilitarian moral judgment. We have found substantial correlations between such scales and AOT (Baron et al., 2015).
4. Measures of criteria for belief, as designed by Metz, Weisberg and Weisberg (2016). These may easily be combined with the last type.
5. Other scales such as those that measure maximization and reflection.

The simulations, in one condition, and the grade-assigning test will both include questions that tap the degree of understanding of AOT as a design.

Third, we shall revise the training used by Gürçay (2016) and examine its effects on measures found to be useful in the first two parts just described.

For many of these studies we can use a panel of about 700 people who have been doing my studies for several years. They have answered questions about AOT before. But we cannot use these for the training studies because too many of them have already done these. And, in addition, we will have to replicate important results on additional samples who may be less familiar with the topics studied. I am resistant to using Mturk, because I like to pay minimum wage, and I fear that this will upset their system. Also, many Mturk workers have done as many similar studies as my panel, so they are not all that different. Thus, I will seek alternative ways of recruiting subjects, and I have put some additional money for this purpose in the proposed budget.

New belief and self-report scale

The first task will be to construct a new version of the self-report scale, building on the current scales when possible. It will include items about direction of both search and search and inference; items about extent of search for possibilities, evidence and goals; and items about decisions and beliefs. (Goals will be examined for decisions, but not beliefs.) For beliefs, it will include items about problem solving (short-term) and long-term beliefs. Leaving goals aside, this amounts to the following topics:

decisions, search for additional possibilities (direction is implied)
decisions, search for additional goals served by current favored option

decisions, search for additional goals not served by current favored option
decisions, search for evidence in favor of current favored option
decisions, search for evidence opposed to current favored option
decisions, extent of search
short-term beliefs, search for evidence in favor of current favored option
short-term beliefs, search for evidence opposed to current favored option
short-term beliefs, extent of search
long-term beliefs, search for evidence in favor of current favored option
long-term beliefs, search for evidence opposed to current favored option
long-term beliefs, extent of search

This amounts to 12 categories. If each had two items we would have a 24-item test. A short version could pick the best item from each category.

The basic form of the test, as now, would be as a questionnaire about the nature of good thinking, that is, how people should ideally think.

Self vs. ideal (for others)

In addition, for experimental purposes. I will modify the items so that they refer to what the subject does rather than what is ideal. This will provide a measure of perceived self-deficiency. But it will also allow me to ask which type of scale is more highly correlated with other measures.

Generality across domains

I will create separate versions of the test for each of several domains. It may be that people think that thinking should differ in different domains. Possible belief domains are: morality, religion, science, current events (news), history, crime (detectives), medicine (diagnosis), aesthetics. Possible decision domains are law (judges), legislation, consumer purchases, personal relationships, personal health decisions, political behavior, personal religious behavior, and academic problem solving. Of course, only the belief items will be used for the belief domains, and only decision items for the decision domains (and also only the short-term or long-term items, as appropriate).

The purpose of these items is to examine overall domain differences and generality of AOT across domains. I will examine the generality of both the ideal and the self-report versions, and their difference. The development of domain-specific versions of the scales is analogous to the method used by Pachur and Spaar (2015), who found lower-than-expected across-domain correlations for a scale that measures intuition and deliberation as cognitive styles. (The domains were all personal decisions: mate-choice, clothing, restaurants, medical, electronics, vacations.)

This procedure will also provide an opportunity to use standard psychometric tools to analyze individual items: measures of reliability and item analysis with item-characteristic curves (Revelle, 2016; Rizopoulos, 2006). These tools are most useful for uni-dimensional constructs. As noted, the theory of AOT does not imply unidimensionalit for the various components of AOT, but it should imply unidimensionality for each component, across domains. So long as we find at least a moderate degree of generality, we can use this prediction to select good items. (However, we should test for

generality before applying this criterion, because selection of items on this basis could increase the observed generality artificially.)

Generality across components of AOT

A 24-item version of the basic scale will allow me to ask about the generality across the various components of AOT. For this purpose the omega() function in the psych package for R (Revelle, 2016) seems useful for a first pass at description. Ultimately, structural equation models may also be useful (available in the sem package).

Simulated decisions and tests of understanding

To test individual differences in AOT, I shall design some simulated decisions in each of several domains: consumer choices, medical decisions, and economic decisions by government. For example, a consumer decision could be simulated by presenting two alternative purchases, as done in many web sites (or, for medical decisions, some symptoms and a proposed diagnosis). The subject would then have a few choices, such as adding more purchases (search for options), searching for positive reviews or negative reviews of one or the other (search for evidence), or possibly reading a short essay about what to look for in products of this type (search for goals). After some steps, the subject would indicate her current evaluation, probably just by indicating which option is favored and how confidence she is that this is the best choice. Each step would produce the relevant information, and then face the subject with another choice.

The design of these simulations would be simple, with only a few steps, but the choices will be designed to asses various forms of myside bias, such as looking excessively for reasons favoring the current leader, or choosing a diagnostic test that will yield a positive result if the favored diagnosis is true but which will fail to distinguish that diagnosis from the next best alternative (Baron, Beattie & Hershey, 1988).

The measures of interest will go beyond selective exposure to supporting evidence. Selective exposure effects are often small and difficult to get. However, individual differences may still exist. And judgments are still affected by prior commitment after the evidence is obtained (Chaxel, Russo & Kerimi, 2013). Moreover, selection of non-diagnostic information (Baron et al., 1988) is not the same bias as selective exposure.

I will probably do the main programming, using a template that can be adapted to any sort of decision. I am a competent JavaScript programmer, as I have been using JavaScript for web studies[3] since 1997. For this purpose I will probably begin with a template[4] written by my ex-student Marianne Promberger. As she explains, I have been generating a frame for each page, and she uses " " tags, with all pages "displayed" at once, except that only one is visible at a time. This permits a simple script to control the ordering of pages as a function of how the subject responds on each one. It is also possible that Qualtrics can do what is needed. I have not been using it, but Burcu Gürçay is highly competent with it.

To test understanding, this procedure will be modified so that the subject does not actually choose what to do but, instead, answers questions about the properties of each choice, such as,

[3] http://finzi.psych.upenn.edu/ baron/ex.

[4] http://promberger.info/template-2.0/.

"This could provide information that would cause me to change my mind," "This is unlikely to cause me to change my mind," "This could confirm my current choice," or "If I do this test and it fails to reject my current idea, then my confidence in that idea should increase." The idea would be to determine how well the subject could distinguish AOT arguments from others. AOT arguments would be correct only for some of the subject's choices on each page.

The use of consumer purchases and medical decisions could be seen as a kind of problem solving in which the subject has no prior commitment. Yet myside bias also occurs in the defense of long-held beliefs, and it is of interest in that context. Previous research using such beliefs (including my own: Baron, 2009) has used controversial issues, where different subjects had opinions on both sides.

I shall develop cases using such issues, which can be put into the framework just described. Possible questions are: whether capital punishment deters homicide; whether raising the minimum wage causes unemployment; whether trade agreements increase unemployment; and whether immigration increases unemployment of natives. These and other questions like them are topical, and it is possible to present subjects with summaries of evidence bearing on both sides.

The trouble with such issues is that many people do not have strong opinions one way or the other. Individual differences in measures of myside bias are confounded by differences in strength of opinions, and such differences in strength of opinion could be general across issues. (People who are high on AOT but also choose to attend to other matters, will correctly lack confidence in their opinions about most such controversial issues.) We will attempt to increase the variety of issues by including other issues that are not sources of polarization but do have arguments against the side that most people believe. These will include historical judgments and conclusions of social and medical sciences. Possible examples are: violence on TV causes aggression; diet soda helps lose weight; the U.S. should spend less on foreign aid; Richard Nixon was a bad president; IQ is not useful as a predictor of success; etc. In addition, we shall look for sources of strong opinion other than politics, such as those that arise from preferences for sports teams or styles of music or apparel.

Grading and analysis of examples

As an alternative way of testing endorsement and understanding of AOT, I will develop questionnaires with several short reports (around 500 words) of someone's thinking, as done by Baron (1995). The subject will assign a grade to each report, and then indicate the reasons for the grade in a yes/no list. The list of reasons will include those related to AOT and built into the examples by design, and, in addition, other reasons that have nothing to do with AOT, such as clarity of expression. The examples will illustrate different errors and will be matched. For example, a group of episodes will consist of a good one, one deficient in search for possibilities, one deficient in search for evidence, and one deficient because of myside bias in inference. Several such groups of items will be used in each session, and the four members of each group will be separated, but otherwise all reports will be presented in a different random order to each subject. The item groups will be chosen from at least two different domains, e.g., science and current events. The tests of understanding would involve correct matching of the reason to the example, so they might not be trivially easy. We shall also include questions about justified confidence, which should be more difficult.

Although this experiment will be long, I will ask the same subjects to do it who have done other studies described above, so that it will be possible to look at correlations. The used of different

domains will also test the generality of the measure, in a different way.

The analysis of responses to the reasons could serve as a test of understanding. If this effort works well, the grading test could become more useful than other methods, given the importance of understanding AOT as well as being able to do it.

Criteria for belief and moral judgment

Metz, Weisberg and Weisberg (2016) asked people about their criteria for belief, with a view to explaining the difference between creationists and those believe the scientific theory of evolution. They found that creationists were more likely to endorse such criteria as: "The Bible says it is true" and "It feels true in my heart." Evolutionists were more likely to endorse "There is good scientific evidence for it" and "It explains a whole lot of things." These results held up when the experiment asked for criteria for belief before mentioning any particular application, such as evolution.

Understanding of AOT, and the role of AOT in science, should increase appreciation for science as a way to arrive at justified beliefs. Thus, we shall use the short questionnaire from the Metz et al. study as a validation criterion.

Training

The current version of the adult AOT training module (described earlier) was designed with the idea of training forecasters in making more accurate predictions, and therefore, is not a general AOT training module. The training was also somewhat specific in the way that the exercises given to participants at the end involved asking questions in the same format as the survey given to them later on (i.e., three multiple choice), so while we could use this training for some of the questions types in the experiments we are planning for the AOT measurement experiments, we need to modify it in order to be applicable for other kinds of question types. Specifically, we will use two different kinds of questions within the training itself and teach for transfer from one to the other, so that the subjects begin to think about transfer (as suggested by other results, beginning with Gick & Holyoak, 1983). Then we will test for transfer using a third type of material, before a post-test of the original two types.

Another shortcoming of the training module was that the subjects were given textbook definitions of vocabulary related to AOT, but they were not made to understand how and why AOT works. We will this give training and testing about the kinds of understanding that we test. Although it is possible that the questions we ask will be too easy, we think they will be more difficult because of the multiple choice format, in which the subject must pick not just any reason why AOT is good but the reason specifically illustrated in each example. This part of the training will be designed after we have some results from the example-grading task described above.

Additionally, even though we observed reduction in the unwarranted confidence, we were not able to test how long the effects of the training would last, or how well they would transfer from the material used in training to other material. We would like to run a study to test the effectiveness of the training module over a longer term. (In the original study, we gave subjects 36 hours to begin the post-test, after doing the training, because we were afraid that effects would dissipate.) It is possible that our subjects might require some shorter refresher modules as they build their AOT skills.

In the future, we would like to develop an online app that would help people develop good thinking skills, and this project would lay the groundwork of our future projects and those of others. We think that this sort of training module, with active refreshers over a period of time, might be something that students and adults might want to do for themselves voluntarily. Schools and other institutions might also encourage it. This is not what we plan to develop in the proposed period, but we should learn better how it might be done.

Broader impacts

Interest in the measurement of AOT seems to be strong at the moment. It is central to the concept of rationality developed by Stanovich, Toplak, and others (Stanovich, 2016; Toplak et al., 2016). The various short scales now in use have been found to be useful predictors in a variety of tasks (e.g., Haran, Ritov & Mellers, 2013; Mellers et al., 2015; Baron et al., 2015). Yet, we have no good theory-based measure of AOT as a trait, and it is possible that some of the correlations of the current scale are distorted (either up or down) by contamination of the scale with other traits. Thus it would be helpful to researchers to have a better scale.

The scales themselves, and the publications that describe them, should provide a clearer definition of AOT and a better explanation of its importance than what is now available. In particular, these products will emphasize the distinction between AOT and other concepts of reflection, the role of understanding AOT in judgments about which sources of information are relatively trustworthy, and, in particular, the relation between AOT and the value of science (as well as other institutions that have traditions that emphasize AOT). Insofar as this clearer understanding of AOT can be "given away" to the culture as a whole, it can improve education and people's thinking in a variety of domains.

Accordingly, in addition to the usual academic outlets, I will try to call attention to these issues (as I already do) through my blog ("Judgment misguided"), other blogs, op-ed pieces written for news outlets, and a fifth edition of "Thinking and deciding" (already underway).

AOT itself is especially important in today's world and merits continued study. The absence of AOT is one of the problems of democratic government (Baron, 2015, in press). AOT is helpful both in making the decisions that citizens need to make and in appreciating the role of science and other disciplines that incorporate AOT into their design. Better measures, especially those that relate to understanding, will allow better research.

Training modules like ours might have practical value in specific contexts, such as training workers in an organization. The impetus for this module in fact came from its possible use for training forecasters in the field of foreign intelligence, and its format was based on similar modules designed for cognitive therapy of depression. However, if training in AOT is to be included in formal education, it will probably require modifications of curriculum in several different areas, in ways that reinforce each other across a student's experiences, as discussed (with examples) by Baron (1993). Beyond this, the culture of the school (if not the culture outside of it) should encourage curiosity and questioning.

Some of the other issues addressed in this proposal will contribute to understanding of how AOT can be increased through education and other means. The studies of generality and training will suggest how much transfer between domains can and does occur.

References

Baron, J. (1985). *Rationality and intelligence*. New York: Cambridge University Press.

Baron, J. (1988). *Thinking and deciding*. New York: Cambridge University Press. (2nd edition, 1994; 3rd edition, 2000; 4th edition 2008).

Baron, J. (1991). Beliefs about thinking. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.), *Informal reasoning and education*, pp. 169–186. Hillsdale, NJ: Erlbaum.

Baron, J. (1993). Why teach thinking? — An essay. (Target article with commentary.) *Applied Psychology: An International Review, 42*, 191–237.

Baron, J. (1995). Myside bias in thinking about abortion. *Thinking and Reasoning, 1*, 221–235.

Baron, J., Badgio, P., & Gaskins, I. W. (1986). Cognitive style and its improvement: A normative approach. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence, Vol. 3*, pp. 173–220. Hillsdale, NJ: Erlbaum.

Baron, J., Beattie, J., & Hershey, J. C. (1988). Heuristics and biases in diagnostic reasoning: II. Congruence, information, and certainty. *Organizational Behavior and Human Decision Processes, 42*, 88–110.

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition, 4*(3), 265–284.

Dewey, J. (1901/2012) *Psychology and social practice*. Orig. University of Chicago Press, Chicago: IL. Reprinted as Project Gutenberg eBook.

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255–278.

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*, 223-241, 263-271

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1–38.

Gürçay-Morris, B. (2016). The use of alternative reasons in probabilistic judgment. PhD Dissertation, Department of Psychology, University of Pennsylvania. http://finzi.psych.upenn.edu/~baron/theses/GurcayMorrisDissertation.pdf (slides for presentation at http://www.sjdm.org/presentations/2016-Talk-Gurcay-Morris-Burcu-alternative-reasons-Probabilistic.pdf).

Haidt, J. (2001). The emotional dog and its rational tale. *Psychological Review, 108*, 814–834.

Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making, 8*, 188–201.

Herek, G. M., Janis, I. L. & Huth, P. (1987). Decision making during international crises. Is quality of process related to outcome? *Journal of Conflict Resolution, 31*, 203–226.

Johnson-Laird, P. N., & Bara, B. G. (1984). Syllogistic inference. *Cognition, 16*, 1–61.

Kagan, J., Pearson, L., & Welch, L. (1966). Conceptual impulsivity and inductive reasoning. *Child Development, 49*, 1005–1023.

Kagan, J., Rosman, B. L., Day, D., Albert, J., & Phillips, W. (1964). Information processing in the child: Significance of analytic and reflective attitudes. *Psychological Monographs, 78* (1, Whole No. 578).

Kahneman, D. (2011). *Thinking, fast and slow.* New York: Farrar, Strauss and Giroux.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 107–118.

Mellers B. A., Stone E., Atanasov P., Roghbaugh N., Metz S. E., Ungar L., & Tetlock P. E. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied, 21*, 1–14.

Messer, S. B. (1976). Reflection-impulsivity: A review. *Psychological Bulletin, 83*, 1026–1052.

Metz, S. E., (2017). Epistemic practices in adults and adolescents. PhD Dissertation, Department of Psychology, University of Pennsylvania.

Metz, S. E., Weisberg, D. S., & Wesiberg, M. (2016). Non-scientific criteria for belief sustain counter-scientific beliefs. Under review.

Mill, J. S. (1859). *On liberty*. London: J. W. Parker & Son.

Pachur, T., & Spaar, M. (2014). Domain-specific preferences for intuition and deliberation. *Journal of Applied Research in Memory and Cognition, 4*, 303–311.

Perkins, D. N. (1986). *Knowledge as design: Critical and creative thinking for teachers and learners*. Hillsdale, NJ: Erlbaum.

Perkins, D., Bushey, B., & Faraday, M. (1986). Learning to reason. Final report, Grant No. NIE–G–83–0028, Project No. 030717. Harvard Graduate School of Education.

Piazza, J., & Landy, J. F. (2013). "Lean not on your own understanding": Belief that morality is founded on divine authority and non-utilitarian moral judgments. *Judgment and Decision Making, 8*, 639–661

Revelle, W. (2016) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.6.9.

Rizopoulos, D. (2006) ltm: An R package for latent variable modelling and item response theory analyses. Journal of Statistical Software, 17(5), 1–25. http://www.jstatsoft.org/v17/i05/

Selz, O. (1935). Versuche zur Hebung des Intelligenzniveaus: Ein Beitrag zur Theorie der Intelligenz und ihrer erziehlichen Beeinflussung. *Zeitschrift für Psychologie, 134*, 236–301.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*, 3–22.

Stanovidh, K. E. (2016). The comprehensive assessment of rational thinking. *Educational Psychologist, 51*(1), 23–34.

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology, 89*(2), 342–357.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General, 127*, 161–188.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Rational thinking and cognitive sophistication: Development, cognitive abilities, and thinking dispositions. *Developmental Psychology, 50*(4), 1037–1048.

Wason, P. C., & Evans, J. St. B. T. (1975). Dual processes in reasoning? *Cognition, 3*, 141–154.

Wertheimer, M. (1959). *Productive thinking* (rev. ed.). New York: Harper & Row (Original work published 1945)