A brief note about post-publication peer-review of this article

The duty of a scientist with a new hypothesis is to:

1. Develop a replicable test of the hypothesis
2. Conduct the test
3. If the test supports the hypothesis, submit a report of it for peer-review
4. If the peers identify a way to improve the test, conduct the improved test; otherwise publish the experiment, so that others may test its replicability

Most peer-reviewers have a specialty and will not endorse work outside that specialty, even if he/she fails to identify a way to improve the test. What if a scientist submits to ten peer-reviewed journals and none is able to find any peer-reviewers who endorse publication or who suggest ways to improve the test? Peer-reviewers are volunteers, so journals cannot force them to expand their specialties. Furthermore, a reviewer who knows no way to improve the test, yet feels uncomfortable endorsing publication, may cite fictional problems or recommend improvements that do not actually change the test procedure. In other words, it is possible for a perfectly legitimate experiment to be blocked from publication merely by problems with the writing/peer-review process.

Post-publication peer-review is appropriate in such cases. If no peer reviewer is able to suggest a change to the testing procedure, the scientist should not abandon the hypothesis. This is a difference between science and politics. This article was written under the guidance of a professor to make sure the writing is clear. It was submitted to ten journals/review systems (sequentially), and received a total of six blind reviews, none of which identified flaws in the test, except to acknowledge flaws common to all survey research.

- If anyone has any difficulty understanding how to replicate this study, please feel free to contact Christopher Santos-Lang directly for further explanation. Amazon's Mechanical Turk makes replication easy and inexpensive. The data is freely available on figshare (https://dx.doi.org/10.6084/m9.figshare.1603027.v1) and the GRIN-SQ is included in this article. Perhaps another scientist who replicates the experiment and will write about it differently and be published in a peer-reviewed journal.
- It is also possible that a reader will identify a way to improve the test. Then it would be the duty of the scientist to conduct the improved test, or at least to acknowledge the need to conduct it. Please document these suggestions via PubPeer.

A history of the peer-review process (including reviewer comments) is included as **Appendix C.**

Measuring evaluative computational differences in humans

Christopher Santos-Lang*

ABSTRACT

Science fiction about intelligent aliens has long imagined a science of sociology with typologies that apply universally, much as the periodic table of the elements applies to atoms on all planets. The GRIN model purports to offer such a universal typology. This study offers the first instrument to measure its manifestation in humans: the Gadfly-Relational-Institutional-Negotiator Self-Quiz (GRINSQ). It reports evidence of the GRINSQ's reliability, as well as its structural, content, convergent and pragmatic validity, including relationships to the Moral Judgment Test (MJT), Moral Foundations Questionnaire (MFQ), and Big Five personality traits. The evidence supports the hypotheses that humans specialize by GRIN-type, and that this specialization relates to differences in personality, morality, political orientation, career, religion, family type, and identification with crime.

1. Introduction

1.1. Interdependent evaluative diversity

Throughout the natural world, entities tend to specialize into universal types, be they types of subatomic particles, elements, cell types, or functional types in an ecosystem. The types are frequently interdependent (e.g. would neuron evolve without muscle? Would uranium come into existence without helium?). One reason to expect interdependent specialization to be advantageous in a society comes from the observation that rate of adaptation is limited by at least four distinct factors:

1. Rate at which novel configurations are produced
2. Selection pressure privileging better configurations
3. Fidelity with which proven configurations are reproduced
4. Network localization

Unless the best approaches for promoting these factors all happen to be the same, the most quickly adapting society will be specialized such that each member promotes only a subset of these factors, yet collectively the members promote them all.

When we construct machines to imitate aspects of society, the four specializations described above manifest as modules in a larger architecture. The tension between the specializations has been labeled "moral disagreement" (as though one in a set of interdependent types could be intrinsically morally wrong), so computer modules were first sorted into these types in the field of machine ethics, where they were called "GRIN-types" [1]:

Gadfly: Unpredictable due to use of novelty (e.g. an individual mutator in evolutionary computation) – compared to pragmatic ethics

Relational: Unpredictable due to network effects (e.g. a cell of level 3 or 4 cellular automata) – compared to virtue ethics

Institutional: Predictably upholds rules (e.g. a standard calculator) – compared to deontological ethics

Negotiator: Predictably converges on maximizing a measurable goal (i.e. supervised machine learning for financial-trading) – compared to consequentialist ethics

Most human brains have regulatory mechanisms, such as creative block, apathy, ego depletion, and learned helplessness, which could prevent the manifestation of a GRIN-type in a given context, thus forcing a human to (temporarily) manifest a different type [2][3][4][5]. In contrast, computers typically rely on human operators to shift the type of software they run. Regulatory mechanisms may allow some humans to switch type like sequential hermaphrodites, but other humans may be less disposed to shift (e.g. Aspies, die-hard conservatives, and highly sensitive persons), and progress would tend to decrease the frequency of switchers over time.

1.2. Computer models

One example of a machine containing gadfly, institutional and negotiator components selects an investment strategy by considering how potential strategies would have performed in former situations [6]. At the highest level, this machine is a negotiator maximizing expected financial returns. However, if the component which generates new potential strategies did not function as a gadfly, the machine would get stuck in the famous problem of local maxima. Furthermore, if the component which implements strategies did not function institutionally, less profitable strategies could be selected over more profitable strategies, so the machine could degrade. Furthermore, the machine is ultimately composed of atoms which function relationally (e.g. subjectively sensitive to nearest-neighbors). Thus, the types in this example are interdependent—the negotiator would not be successful as a negotiator if its gadfly, institutional, and relational components did not function in their non-negotiator ways.

Not all computers function as negotiators at the highest level. The first popular computers functioned institutionally at the highest level. Machines that behave unpredictably (i.e. as gadflies or relationally at the highest level) are not currently popular beyond entertainment and the laboratory, but they do exist.

The gadflies used in modern computers tend to be simple random number generators, but human gadflies may have more sophisticated means to avoid predictable paths. For example, they may exhibit attraction to unsolved problems, rebellion, or the "cutting-edge." Likewise, although both humans and computers are able to serve relational, institutional and negotiator functions, rarely are a human's emotional bonds, rules, or goals, respectively controlled by a "programmer" (Milgram [7] is an exception). Thus, although we may call ourselves "slaves" to our emotional bonds, rules, or goals in some sense, humans typically do not have masters in the same sense modern computers do. We could specialize into GRIN-types without being exactly like computers.

* Tel: +1 920 747 0335, *E-mail address:* chris@GRINfree.com

1.3. Evaluativism

The hypothesis that humans already divide by GRIN-type yields predictions for both psychology and sociology. The current study focuses on testing predictions about psychology, but would be remiss not to consider whether the hypothesis has already been falsified at the sociological level. To specialize by GRIN-type would create pervasive fundamental biases in individual humans. Thus, we would disagree with people of other types, have difficulty understanding why they disagree with us, and have difficulty recognizing our interdependence. If humans already divide by GRIN-type, we should be able to observe discrimination on the basis of GRIN-type, and reduced rate of adaption wherever that discrimination is allowed to bloom into segregation.

Consistent with these predictions, at least the first century of psychological research into disagreement found types of people, but assumed their disagreement stems purely from differences of error, illness, and immaturity. It discriminated by ranking the types, never exploring the possibility that some types might be interdependent [8]. Much as the earliest philosophies of racism claimed to justify discrimination against people who do not share one's race, recent philosophies of evaluativism claim to justify writing-off anyone who does not share one's own evaluations (e.g. [9]). Political polarization is a familiar example. Measurements show that evaluativism currently outpaces racism [10][11] as a form of discrimination, and that evaluativism is so severe in the average home that children's self-reported values tend not to align with their own genetic predispositions until they leave their parents [12].

In addition to predicting the existence of evaluativism, the hypothesis that humans already specialize by GRIN-type predicts that evaluativism would handicap societies much as speciesism can handicap ecosystems (i.e. by leading to the disabling of components upon which the system as a whole depends [13]). For practical and ethical reasons, it is rare to conduct controlled experiments manipulating levels of evaluativism in societies, but those which have been conducted indicate that competitive design teams win only half as much when evaluativism is allowed to run rampant within them [14][15].

1.4. Current study

Confirmation that societies inevitably specialize into universal interdependent types would be expected to include validation of an instrument which can sort members of our own society into those types, and the instrument would be expected to yield bimodal distributions (a serious criticism of the MBTI and other attempts to measure personality types). While cluster analysis (e.g. [16]) and behavioral measures (e.g. [17][18]) imply division into distinct types, inability to develop the expected survey instrument muddies these results. Existing survey instruments measure traits instead of types (e.g. [19][20][21]), or proximity to a single type (e.g. [22][23][24]). To the best of our knowledge, the current study is the first to fulfill the sorting expectation, offering a survey instrument that produces the required distributions, and providing evidence of its reliability and structural, content, convergent, and pragmatic validity.

1.5. Selection of Assessment Type

There were several reasons to choose forced-choice questions over Likert-type scales. The major advantage of Likert-type questions is to produce scalar numbers permitting analysis via correlation, factor-analysis, and regression [25], but this advantage is illusory with categorical constructs like GRIN-type and species. No matter how scalar one's measure of "humanness," for example, it would be invalid to call one person "more" human than another or to perform regressions which suggest strategies to become "more" human. Analyzing GRIN-type in degrees would be just as absurd.

Likert-type questions also have some important disadvantages. The first is that they support an illusion that one's construct is complete. A classifier based on forced-choice questions (e.g., a species classifier) leaves some subjects unidentified, thus exposing the incompleteness of the construct, and leaving a path to improve it.

The second major disadvantage of Likert-type scales is their inability to distinguish subjects with more reliable results. If subjects hide their types (which we would expect, based on the sociological predictions), they might not do so equally. Forced-choice batteries allow us to assess reliability on a subject-by-subject basis, naturally classifying subjects with unreliable results as unidentified. Likert-type batteries, in contrast, misrepresent random answers as valid (i.e. falsely indicating balance).

1.6. Development of the GRINSQ

The GRIN Self-Quiz (GRINSQ) in **Appendix A** was developed through iterative rounds of item generation and selection. We started with four-way forced-choice questions with one choice per GRIN type. Respondents ranked the choices, effectively expressing six pair-wise comparisons per question. The goals of revision were to maintain content validity while maximizing internal consistency within, and discrimination across, all comparisons. Data was collected through verbal protocols, online survey administration, and administration in high-school classrooms. Different types of questions are more relevant for different comparisons, so only the most discriminating four sets of six pair-wise comparisons (24 comparisons total) were retained in the final version of the GRINSQ.

2. Material and methods

2.1. Participants

For validation, an independent standard U.S. sample of 250 participants was recruited by Amazon's Mechanical Turk. As a trusted third-party, Amazon pays the "Turkers," maintains records of their consent, gives them freedom to choose which surveys to complete (if any), and affords them anonymity with respect to investigators. This was a typical survey on a single page in Mechanical Turk with no deception or experimental manipulation, and it was made clear that the data would be used for publically published research, so no additional consent was gathered. It is typical for Institutional Review Boards (IRBs) to exempt such protocols from review. Each subject was paid $0.50

to answer a total of 124 questions, including the GRINSQ, BFI-10, MJT, MFQ, and a battery of demographic questions. Numerous studies have confirmed that data gathered through Mechanical Turk is at least as reliable as that gathered through traditional methods, even when compensation is low [26][27[28].

An additional nine Subject Matter Experts (SMEs) were recruited via AI-themed listservs to assess the content validity of the GRINSQ via an anonymous online survey. Each reported having an advanced degree in computer science/philosophy and/or substantial publication/work experience with artificial intelligence.

2.2. Materials and Procedure

The MFQ used to measure endorsement of the moral intuitions of harm, fairness, authority, loyalty and sanctity was developed by Graham, Haidt & Nosek [29]. It is currently a standard instrument of moral psychology. To measure the Big Five personality traits of openness, conscientiousness, extraversion, agreeableness, and neuroticism, we used the 11-item BFI-10 which Rammstedt & John [30] demonstrated has good psychometric properties despite its brevity.

The abridged Moral Judgment Test (MJT) by Lind [31] used in this study consisted of two sets of scales: one regarding the Doctor's Dilemma (euthanasia), and the other regarding the Worker's Dilemma (procedural justice). The MJT is typically used for its c-index, which measures how consistently a subject ranks moral arguments of different types, but was selected for this study because it also offers factors for each of Kolberg's six developmental stages of moral reasoning [32]. The other standard survey instruments for assessing moral reasoning, the DIT and DIT2, do not offer stage-wise factors—they measure only proximity to a privileged type [33][34].

The online survey administered to SMEs asked each to list his/her credentials, to give examples of machine types beyond the GRIN model, and then to rate each content assumption of the GRINSQ on a 5-point Likert scale (from "Disagree strongly" to "Agree strongly"), assuming the following meanings for non-technical terms used in GRINSQ items:

1. The words "person" and "feeling" are interpreted to include machines and their internal states.
2. The words "empathy," "love," and "relationship" are interpreted as referring to participation in network effects (i.e. responding differently to entities closer to oneself).

Each content assumption was phrased as a hypothesis about how a machine would answer a GRINSQ item, and why. For example, SMEs were asked to rate their agreement that a mutator (gadfly) would answer the first question as (B), "My higher priority in life is to discover new possibilities" because "mutation is useless if unable to produce anything new."

3. Results and Discussion

3.1. Sample Characteristics

Typical of Mechanical Turk, the 250-person sample was biased towards liberalism (57%) and young adult ages; only 98 subjects were over 35, and none were under 18 [35]. The sample was otherwise balanced, as shown in Table 1.

Table 1: Characteristics of the 250-person sample

N	*%*	*Characteristic*
120	48%	Male
130	52%	Female
58	23%	Age 18-24
94	38%	Age 25-34
50	20%	Age 35-49
48	19%	Age 50+
187	75%	White
20	8%	Asian
18	7%	Hispanic
13	5%	Black
24	10%	Disabled/handicapped
126	50%	Completed college
124	50%	Not yet completed college
57	23%	High socioeconomic status
138	55%	Medium socioeconomic status
55	22%	Low socioeconomic status
48	19%	Resident of major city
58	23%	Resident of minor city
98	39%	Resident of suburb
46	18%	Resident of town/country
107	43%	Christian
104	42%	No religion
39	16%	Non-Christian religion
142	57%	Liberal
59	24%	Moderate
49	20%	Conservative
122	49%	Have been in a committed relationship (e.g. marriage)
84	34%	Have served as a parent
73	29%	Have served as a manager
52	21%	Have started a business/nonprofit
44	18%	Have served the disadvantaged (e.g. mission work)
41	16%	Have been accused of a crime/serious betrayal

3.2. Content validity

None of the nine SMEs was able to offer an example of a machine beyond the GRIN model. One warned that predictability may be less a property of a machine than of the predictor (though this did not prevent classification in practice). Four volunteered that the same hardware could run multiple GRIN types.

Table 2 shows mean SME agreement for each assumption of the GRINSQ. On a scale of 1-5 where 3 is "neither agree nor disagree," average SMEs agreement was 3.5 for institutional, 3.7 for relational, and 3.9 for both gadfly and negotiator. The table is organized around the properties of GRIN machines as follows:

- To be useful, ***gadflies*** (e.g. mutators in evolutionary computation) need to be part of a larger system to which they propose changes (4A, 7A, 13B). As input, they need access to a randomness generator or other source of novelty (19A) which they use to generate alternatives to existing strategies (16A, 24A). As output, gadflies excel at invention (1B, 10B, 12A, 18B) and escape from local maxima (6B, 22B).
- To be useful, a ***relational*** machine (e.g. a cell in level 3 or 4 cellular automata) needs to be part of a network of other machines (16B, 20B, 23B) which influence and are influenced by it (4B, 5B, 14A). Network effects emerge because relational machines are more sensitive to closest-relations (2A, 11A, 22A).

Complexity may emerge from the network as a whole, but the output of each individual relational machine is merely to preserve diversity through localization (8B, 10A, 17B).

- ***Institutional*** machines (e.g. a standard calculator) work better on input for which random noise is bounded (9A, 14B, 21A). They function by implementing predefined objective rules (20A, 24B). Their output is consistent and exact (2B, 6A, 15B, 18A). This makes them exceptionally qualified to preserve institutions (3B, 8A, 12B).
- As input, ***negotiators*** (e.g. supervised learning machines) require goals (1A, 3A, 17A) and often benefit from seed strategies (13B, 23A). They function by shifting to whichever strategy produces the most success (7B, 9B, 19B), which they discover via diverse mechanisms, often leveraging sub-components of the other types. Their output is convergence toward their goal (5A, 11B, 15A, 21B).

Via open-ended comment, six of the nine SMEs expressed serious caution about the assumptions required to generate these ratings, asserting that personhood goes beyond mechanical phenomena. This warning could stem from threat to human egos. Consistent with this hypothesis, SMEs agreed least with the comparison most threatening to human egos (i.e. that between humans and standard calculators).

The SMEs did not wholeheartedly endorse the GRINSQ, but the GRINSQ did survive the test of content validity. If the GRINSQ measured something other than proclivities for GRIN algorithm types, we would expect SMEs to agree about which items violate that intent. They had no such agreement.

Table 2: Subject Matter Expert (SME) agreement that GRIN exemplars would answer as assumed

Item(s)	*Content assumption (text and rationale)*	*Mean Agreement*
Mutator (gadfly) Input		
19A	I am more concerned about stress which blocks my creativity. *Rationale:* because mutation is useless if unable to produce anything new.	3.9
4A, 7A	I cannot be my best self when my work does not require creativity. *Rationale:* because the relative inefficiencies of mutation are justified by its potential for greater creativity.	3.8
13B	I would prefer to have no plan than to have unquestioned authorities. *Rationale:* because mutation is a way of questioning authorities.	3.7
Mechanics		
16A, 24A	More than others do, I question existing best practices. *Rationale:* because the purpose of mutation is to seek alternatives to the status quo.	3.9
Output		
6B	People who have known me longest treat me as if I am more idealistic but impractical. *Rationale:* because most mutations do not improve upon the status quo.	3.8
18B	People are more likely to complain about my far-fetched proposals. *Rationale:* because most mutations do not improve upon the status quo.	3.9
22B	In the future, I will convince others to open their minds. *Rationale:* because the function of mutation is to raise awareness of new possibilities.	3.9
1B, 10B, 12A	My higher priority in life is to discover new possibilities. *Rationale:* because mutation is useless if unable to produce anything new.	4.0
Cellular atomaton (relational) Input		
4B	I cannot be my best self when I cannot empathize. *Rationale:* because the function of cellular automata relies on network effects (adapting its state according to those of close entities).	3.8
23B	I would prefer to have no plan than to have a "pure business" culture. *Rationale:* because a "pure business" culture demands objective logic (not accommodating special relationships), which is the opposite of the way individual cellular automata work.	3.4
16B, 20B	More than others do, I maintain relationships. *Rationale:* because being positioned in a network (i.e. having relationships) is part of the definition of cellular automata.	3.6
5B, 14A	My higher priority in life is to be lovable. *Rationale:* because the function of a cellular automaton relies on its ability to participate in network effects.	3.7
Mechanics		
2A, 11A, 22A	In the future, I will focus on the people I love the most. *Rationale:* because a cellular automaton prioritizes the states of entities adjacent in its network.	3.7
Output		
8B, 10A, 17B	My higher priority in life is the feelings of the people closest to me. *Rationale:* because a cellular automaton prioritizes the states (including feelings) of other entities closest to it.	3.7

3.3. Reliability/Internal Consistency

Reliability of the GRINSQ is measured per respondent—because of its structure, a significant score (i.e. above cut-off) cannot be obtained without internal consistency. In this sample, 73% had significant scores (*see Structural Validity*).

For Likert scales, reliability would be measured in terms of alpha as reported in **Table 3**. Alphas for the forced-choice questions are handicapped relative to Likert scales because many of the pair-wise comparisons offer no choice relevant to the subject's type, thus alphas for the GRINSQ are not directly comparable to those for Likert scales (though they would be comparable to those of future versions of the GRINSQ). A 0.69 average was measured for the GRINSQ (the four components of each GRIN factor are scores for a quarter of the quiz), which is excellent considering the handicap.

The alphas for MJT stage factors were very poor (ranging from -0.02 to 0.26), but that might be typical for the MJT. Previous studies of the MJT assert its validity on the basis of stronger correlations between adjacent stages, preference for higher stages, and correlation between preference for higher stages vs. consistency of stage preference [36]. MJT stage factors were dropped from the remainder of the study because they do not meet standard tests of reliability. Discounting the MJT makes the GRINSQ the first survey instrument to distinguish more than three evaluative types in humans.

3.4. Structural Validity

The GRINSQ purports to discern the four GRIN types, so its overall factors should be distinct yet correlate internally. Correlations are not standard statistics to describe relationships among types—correlations are given in **Table 4** only to show general uniformity of components. The test of structural validity for the GRINSQ is in the bimodal distributions of its factors. **Fig 1** shows the raw distributions for all four factors along with what would be expected if subjects answered at random. The expected bimodal distribution is seen in their exceeding the random

Table 2 ***(cont.)***

Item(s)	*Content assumption (text and rationale)*	*Mean Agreement*
Standard calculator (institutional) Input		
9A	I am more concerned about stress which leads me to experiment with less-pure behaviors. *Rationale:* because the value of a calculator depends upon behaving consistently.	3.7
14B, 21A	My higher priority in life is to exercise self-discipline. *Rationale:* because the function of a calculator relies on maintaining consistent behavior.	3.9
Mechanics		
20A, 24B	More than others do, I uphold moral principles. *Rationale:* because the function of a calculator is to apply the principles of arithmetic consistently (as if it considered them to be moral principles).	3.0
Output		
2B, 15B	In the future, I will stay pure. *Rationale:* because the value of a calculator depends upon staying consistent.	3.8
18A	People are more likely to complain about my old-fashioned morals. *Rationale:* because staying consistent will eventually make a calculator old-fashioned and it clings to its rules as though it considered them to be moral.	3.3
6A	People who have known me longest treat me as if I am more morally strict. *Rationale:* because a standard calculator follows the rules it has been given (of arithmetic) strictly, as though it considered them to be moral.	3.8
3B, 8A, 12B	My higher priority in life is to serve something greater than myself. *Rationale:* because a standard calculator serves the rules of arithmetic without question (and is built to do so because those rules are esteemed).	3.1
Supervised learning machine (negotiator) Input		
1A, 3A, 17A	My higher priority in life is to know what I am trying to achieve. *Rationale:* because supervised learning must be structured around a goal.	3.8
13B, 23A	I would prefer to have unquestioned authorities or a "pure business" culture than to have no plan. *Rationale:* because supervised learning progresses by improving its plan for how to respond to varying input, so lacking a plan would mean starting over from scratch.	3.6
Mechanics		
7B	I cannot be my best self when I do not know the criteria by which success is measured. *Rationale:* because measurements of success are necessary for supervised learning.	4.1
9B, 19B	I am more concerned about stress which puts my plans on hold. *Rationale:* because a learning machine can have a timely goal, so being put on hold can be a concern.	4.0
Output		
11B, 15A	In the future, I will make measureable achievements. *Rationale:* because supervised learning requires making achievements which are measurable.	4.0
5A, 21B	My higher priority in life is to get results. *Rationale:* because the priority of a supervised learning machine is to maximize on its goal.	3.7

Table3: Cronbach's alpha by subsample

		Cronbach's alpha								
Factor	*Items*	*TOTAL*	*MALE*	*FEMALE*	*18-24*	*25-34*	*35-49*	*50+*	*COLLEGE*	*NO COLLEGE*
		N=250	N=120	N=130	N=58	N=94	N=50	N=48	N=126	N=124
GRINSQ (avg)		0.69	0.64	0.73	0.70	0.65	0.65	0.72	0.70	0.68
Gadfly	4	0.72	0.66	0.75	0.70	0.73	0.72	0.70	0.73	0.71
Relational	4	0.69	0.62	0.73	0.59	0.64	0.80	0.74	0.71	0.68
Institutional	4	0.74	0.66	0.78	0.81	0.67	0.68	0.70	0.70	0.77
Negotiator	4	0.62	0.60	0.65	0.70	0.54	0.39	0.73	0.67	0.57
MFQ (avg)		0.71	0.67	0.72	0.66	0.71	0.73	0.72	0.70	0.71
Authority/subversion	6	0.72	0.68	0.75	0.72	0.74	0.72	0.71	0.72	0.72
Care/harm	6	0.63	0.53	0.62	0.50	0.60	0.68	0.65	0.62	0.63
Fairness/cheating	6	0.62	0.62	0.61	0.51	0.58	0.71	0.67	0.60	0.63
Loyalty/betrayal	6	0.75	0.76	0.75	0.75	0.80	0.73	0.73	0.76	0.74
Sanctity/degradation	6	0.82	0.74	0.87	0.83	0.82	0.84	0.82	0.82	0.82
BFI-10 (Avg)		0.62	0.58	0.65	0.53	0.58	0.70	0.64	0.62	0.61
Agreeableness	3	0.57	0.54	0.59	0.49	0.54	0.66	0.62	0.64	0.49
Conscientiousness	2	0.59	0.57	0.60	0.50	0.60	0.62	0.62	0.55	0.61
Extroversion	2	0.64	0.56	0.70	0.71	0.36	0.73	0.79	0.64	0.65
Neuroticism	2	0.79	0.75	0.81	0.73	0.79	0.85	0.78	0.79	0.79
Openness	2	0.50	0.46	0.53	0.22	0.60	0.63	0.41	0.50	0.51
MJT (avg)		0.14	0.08	0.16	0.03	0.08	0.21	0.13	0.11	0.19
Stage 1	4	-0.02	0.15	-0.26	-0.12	0.17	-0.14	-0.25	-0.14	0.10
Stage 2	4	0.20	0.06	0.32	0.24	-0.36	0.25	0.64	0.22	0.30
Stage 3	4	0.09	-0.15	0.26	-0.05	0.05	0.23	0.18	0.06	0.16
Stage 4	4	0.19	0.21	0.18	-0.13	0.24	0.45	0.18	0.22	0.13
Stage 5	4	0.26	0.26	0.26	0.33	0.20	0.46	0.00	0.32	0.22

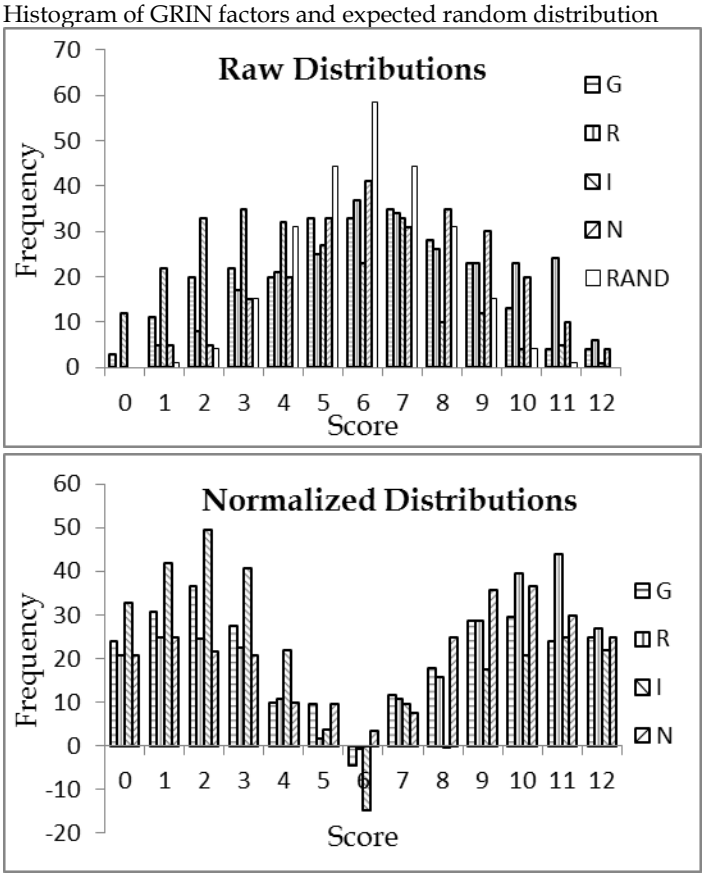
Table 4: Correlations of factors and components

			Correlation			
Component	*Mean*	*S.D.*	*G*	*R*	*I*	*N*
G1	0.06	0.93	**0.761**	-0.268	-0.366	-0.137
G2	-0.18	0.94	**0.744**	-0.300	-0.409	-0.035
G3	0.00	0.87	**0.738**	-0.347	-0.345	-0.048
G4	-0.03	1.03	**0.715**	-0.217	-0.347	-0.163
R1	0.27	0.92	-0.266	**0.750**	-0.245	-0.270
R2	0.16	0.95	-0.372	**0.822**	-0.155	-0.333
R3	0.38	0.78	-0.384	**0.778**	-0.093	-0.339
R4	0.10	1.09	-0.117	**0.658**	-0.161	-0.427
I1	-0.44	0.91	-0.414	-0.142	**0.823**	-0.299
I2	-0.38	0.94	-0.331	-0.209	**0.740**	-0.225
I3	-0.50	0.84	-0.395	-0.124	**0.751**	-0.260
I4	-0.16	0.95	-0.368	-0.187	**0.718**	-0.183
N1	0.11	0.81	-0.065	-0.330	-0.264	**0.734**
N2	0.41	0.78	-0.057	-0.419	-0.147	**0.695**
N3	0.12	0.89	0.001	-0.334	-0.303	**0.709**
N4	0.09	0.91	-0.231	-0.215	-0.205	**0.724**

distribution at the extremes. This is clarified in **Fig 2**, where the distributions are normalized by subtracting the random distribution.

One advantage of polymorphic constructs is that they allow validity to be measured on a subject-by-subject basis. For example, it is possible to know that one has discovered a new species because the species classifier fails measurably on the relevant specimen. The probability of randomly generating GRIN scores with none greater than eight is over 50%, so such results are measurably ambiguous. In our sample, only 66 subjects (27%)

Figures 1 and 2:
Histogram of GRIN factors and expected random distribution



had such a result. This doesn't necessarily prove that all 66 had type(s) beyond GRIN—some might have answered somewhat randomly, or misunderstood questions, or misunderstood themselves—but these 66 would be a good place to start the search for new types.

For the remaining 184 (73%), the GRINSQ produced evidence that could at least supplement evidence from other sources. A score greater than ten would be considered statistically significant by itself ($p<0.015$) because the probability of generating a set of scores with at least one greater than ten is only 1.5%. Our sample included 58 subjects (23%) with such scores: 8 gadfly, 30 relational, 6 institutional, and 14 negotiator. Collectively, they confirm that at least four distinct orientations can be discerned among Mechanical Turkers in the United States. The relative frequencies in this sample likely reflect the fact that liberals outnumbered conservatives nearly three to one.

3.5. Convergent/Divergent Validity

The GRINSQ is the first survey instrument to measure computational evaluative differences, so there is no other instrument with which we should expect one-to-one correspondence. We expect significant relationships with the MFQ and BFI-10 to the extent they measure differences in the ways people evaluate, but the MFQ and BFI-10 measure traits rather than types, so we tested these relationships via t-tests on subsamples with GRINSQ scores above eight, as shown in **Table 5** (equivalent results were found when the cut-off was raised from eight to nine). Results on the GRINSQ may shift for some humans over time, but significant relationship of the GRINSQ to these stable traits implies that GRINSQ results will have similar stability for many people.

Table 5:
Cohen's d for subsamples with factor scores above a cut-off

	Subsample			
Measure	*G > 8* (N=44)	*R > 8* (N=76)	*I > 8* (N=22)	*N > 8* (N=64)
BIG FIVE				
Agreeableness	-0.13	**0.55****	0.19	***-0.41****
Conscientiousness	-0.13	0.11	0.22	0.08
Extraversion	0.35	-0.05	-0.30	-0.23
Neuroticism	-0.22	0.09	0.00	0.13
Openness	**0.50****	-0.07	-0.05	-0.06
MFQ				
Authority/subversion	***-0.78****	0.09	**0.69****	-0.14
Care/harm	0.04	**0.38****	-0.04	-0.23
Fairness/cheating	0.12	0.13	-0.29	0.02
Loyalty/betrayal	***-0.69****	0.08	0.28	-0.17
Sanctity/degradation	***-0.78****	-0.04	**1.25****	-0.24
MJT				
C-index	***-0.20***	0.01	0.02	0.01
Worker's dilemma	0.19	-0.21	-0.34	-0.06
Doctor's dilemma	-0.31	0.13	***-0.87****	0.18

** $p<0.01$

All expected relationships were exhibited and no unexplainable relationships were found: Subjects identified as gadflies exhibited significantly higher scores for openness ($p=0.0031$), and lower endorsement of the moral intuitions of authority ($p<0.0001$), loyalty ($p=0.0001$), and sanctity ($p<0.0001$). Openness is expected for gadflies because gadfly algorithms are designed to be open to a wider set of possible solutions. Subjects identified as gadflies also exhibited a lower c-index ($p=0.0745$) which was expected because gadfly algorithms are designed to be unpredictable.

Subjects identified as relational exhibited significantly greater agreeableness ($p<0.0001$) and endorsement of care ($p=0.0049$). Agreeableness and endorsement of care are expected for relational subjects because relational algorithm components derive their values from closest relations, and care is associated with empathy [37], which is one of the ways humans derive values from closest relations.

Subjects identified as institutional exhibited significantly higher endorsement for authority ($p=0.0063$) and sanctity ($p<0.0001$) and higher condemnation of euthanasia in the Doctor's Dilemma on the MJT ($p=0.0005$). Endorsement of authority and sanctity (i.e. purity) are expected from institutional subjects because institutional algorithms exhibit the purest obedience.

Finally, subjects identified as negotiators exhibited average endorsement of moral intuitions but significantly lower agreeableness ($p=0.0027$). Negative agreeableness (i.e. competitiveness) is expected from negotiators because negotiator algorithms function via competition.

3.6. Pragmatic Validity

Graham et al. [38] argued that new scales merit attention only if they allow us to support important new conclusions. For example, they demonstrated the pragmatic validity of the MFQ by showing that it allows scientists to explain the intractability of political disagreement as stemming from moral differences. Likewise, the GRINSQ supports new explanations of various phenomena. These explanations are supported through chi-squared test as tabulated in **Appendix B**.

Like the MFQ, the GRINSQ shows significant relationship to political orientation ($p<0.0001$). Subjects identified as gadflies were nearly four times as prevalent among liberals as among conservatives, and those identified as institutional were nearly ten times as prevalent in the opposite direction, but negotiator was the type most prevalent among those who consider civics/politics part of their identity (41%). Negotiator algorithms win competitions, so it makes sense that negotiators would rise in modern politics. The MFQ does not distinguish negotiators, so the GRINSQ allows a deeper explanation for political deadlock: that negotiators control politics and find advantage in focusing debate on what divides the other types.

Second, the GRINSQ supports new explanations of vocation/avocation. Categorizing subjects' self-reported careers according to the Holland typology [39] reveals that "artistic" careers were about twice as likely to be occupied by subjects identified as gadflies as by subjects of any other type ($p=0.0023$). Subjects who considered child-care and romance part of their identity were about twice as likely to be identified as relational as anything else, and subjects identified as negotiators were significantly less likely to remain financially dependent on their families past the age of twenty-five, and significantly more likely to occupy "enterprising" careers (i.e. influencing others, as in business and politics). Subjects who identified with team sports were three times as likely to identify as a negotiator than as any

other type. This evidence suggests that society evolved discrete careers and hobbies to match the discrete GRIN-types. Instead of a little angel on one shoulder and a little devil on the other, the cartoonist might more accurately depict our inner-life as divided between an artist-self, family-self, student-self, and conqueror-self.

These results may make one wonder about the morality of circumstances which force a person into a particular career or avocational activity, including forced participation in family, school, competition, and religion. Respondents who converted to no religion were about ten times as likely to identify as gadflies or negotiators as those who converted to Christianity. Because Christianity was such a dominant religion in our sample, it is difficult to tell whether this relationship is specific to Christianity or to whichever religion happens to dominate the sampled region. Major Christian teachings about GRIN orientations are replicated in each of the six other most common religions of the world [40], so it is doubtful that Christian doctrine has any special GRIN bias. However, the evidence demonstrates a profound GRIN-type bias in the average U.S. religious social environment.

The third phenomenon for which the GRINSQ supports new explanation is patterns of distrust. Among the 41 subjects who reported being accused of a crime or other serious betrayal of trust, 29% and 27% identified as gadflies or negotiators, while only 15% and 2% identified as relational or institutional. One might hypothesize that gadflies and negotiators are more likely to be guilty, but that would beg the question. Should we expect the same behavior from people who evaluate differently? If we were measuring scalar personality traits such as psychopathy, we might write-off gadflies and negotiators as "outliers," but our construct is a set of types. Computer scientists cannot solve the widest range of problems efficiently and reliably without all four types in their algorithmic toolboxes, so none can be written-off. The results in this study raise the possibility that our society might expand the range of puzzles we can solve by managing trust and peacekeeping differently.

The evidence seems especially strong when we account for age. Relational orientation is more frequent after the age at which one typically starts a family (rising from 21% to 31%), and institutional orientation is more frequent after the age at which one typically becomes dependent on institutions (rising from 4% to 19%). Thus, a priori, relational and institutional subjects would be expected to report having faced more accusations because they tend to be older and therefore to have had more opportunities to be accused. On the other hand, perhaps the reason why we find older subjects less likely to report negotiator behaviors and attitudes (dropping from 26% to 10%) is that older negotiators have adapted to the pressure of evaluativism by habitually hiding their evaluative identities (even on an anonymous survey).

4. Conclusions

The primary research question for this study was whether it is possible to discern GRIN-types in humans. We were able to develop a reliable self-quiz, the GRINSQ, which has the expected structure, scope of content, and relationships with the BFI-10 and MFQ. This supports the conclusion that GRIN-types exist among humans, and is a step towards establishing GRIN-type as a universal social typology.

4.1. Research Limitations

Two different kinds of inference were used to reach the conclusions of this study:

1. *Induction* was used to conclude that the GRINSQ is reliable and has structure and relationships to the BFI-10 and MFQ consistent with being a measure of GRIN-type.
2. *Abduction*, also called "inference to the best explanation," was used to conclude that whatever the GRINSQ measures actually is GRIN-type.

Both conclusions are valid, given the evidence to date, but, because they employ different kinds of inference, different events would be required to invalidate them. To invalidate the inductive conclusions would require gathering data that exhibited contrary patterns. The provided calculations of statistical significance indicate how unlikely such data is to arise. On the other hand, to invalidate the abductive conclusion would merely require devising a better explanation for the results of the validity tests. The inductive proof in this study passed a remarkably high bar, but the bar for the abductive proof was inevitably low, since very few scientists have yet had opportunity to devise alternate explanations for the results of the validity tests.

Entertaining the notion that GRIN-types do not exist, three alternate explanations for the results of the validity test might be that the GRINSQ instead measures differences of morality or of personality or of politics. Mounting such explanations would involve reworking our understanding of morality, personality, or politics towards better fit with the GRIN model, thus producing many of the same practical results as the conclusion that GRIN-types exist (i.e. type structure, mapping to computer science, etc.). Our inability to know whether future scientists will choose the label "GRIN-type" or "moral type" or "political orientation" or "personality" or something else entirely is a consequence of the current state of social science.

Finally, self-report measures of type can be problematic because of their potential to produce social privilege. For example, in a society of one thousand women and only one man, the man may enjoy special social power by virtue of his gender. Likewise, if society creates (or has created) privileged think-tank positions to support the functioning of natural gadflies, many people may pretend to be natural gadflies so as to secure those positions. Similarly, prison inmates might disguise their GRIN-types if they thought it would help to reduce their sentences. Because of such issues, society needs to develop measures which do not rely on self-report. The GRINSQ may play a crucial role in the validation of those instruments, much as the BFI-10 and MFQ played crucial roles in the validation of the GRINSQ.

4.2 Directions for Future Research

Specialization into interdependent types may already be our best explanation for why GRIN-types appear to exist, but this hypothesis (and the implication that we have reason to moderate evaluativism) can and should be tested in additional ways. Wilde demonstrated interdependence among specializations by comparing the successes of teams with different diversity mixes [41][42]. Including the GRINSQ as an additional measure of diversity in such experiments could provide evidence about whether GRIN-types are interdependent. Such studies could even test whether suppression of human gadflies reduces rate at which novelty is produced, whether suppression of relational humans leads to less localized social networks, whether

suppression of institutional humans leads to more reinvention of the wheel, and whether suppression of human negotiators reduces the accuracy of tests of novel configurations.

Validation of the GRINSQ through comparison to other survey instruments may justify investment in comparisons to non-survey measures (e.g. [43][44][45]). If we think evaluativism blinded previous psychologists, causing them to conceive individual differences in terms of rank (i.e. the moral people vs. the immoral people), it may make sense to measure overlap of the GRINSQ with constructs like IQ and psychopathy which were also conceived in terms of rank.

The GRINSQ may also help us understand discrimination. Evaluativism has been shown to be implicit [46], so it may have become institutionalized without our realizing it. For example, there may be specific church traditions, industry practices, and even laws which discriminate against people of particular GRIN-types. The current study found that people of certain GRIN-types are more likely to be accused of a crime or other serious betrayal of trust, but we should like to know which kinds of crimes or betrayals those would be. Like the victims of homophobia, the victims of evaluativism can turn-out to be closer to us than we expected. We should like to know whether the institutions we design and preserve for our grandchildren to inherit are likely to oppress them. Monitoring the impact of specific social practices on the manifestation of specific GRIN-types could help us answer that question.

Acknowledgements

Thanks to all the editors and reviewers who provided feedback on earlier drafts of this manuscript. Special thanks to the hundreds of Turkers, students, family, and friends who served as guinea pigs to develop and validate the GRINSQ.

References

1. Santos-Lang C. Moral ecology approaches to machine ethics. In: van Rysewyk S, Pontier M, editors. *Machine Medical Ethics*. New York: Springer; 2014. pp. 111-127.
2. Arias–Carrión O, Pöppel E. Dopamine, learning and reward–seeking behavior. *Acta Neurobioligiae Experimentalis*. 2007; 67: 481–488.
3. Grace AA, Floresco SB, Goto Y, Lodge DJ. Regulation of firing of dopaminergic neurons and control of goal-directed behaviors. *Trends in neurosciences*. 2007; 30(5): 220-227.
4. Hagger MS, Wood C, Stiff C, Chatzisarantis NL. Ego depletion and the strength model of self-control: a meta-analysis. *Psychological bulletin*. 2010; 136(4): 495-525.
5. Zak P. The physiology of moral sentiments. *Journal of Economic Behavior and Organization*. 2011; 77: 53–65.
6. Santos-Lang C. Our responsibility to manage evaluative diversity. *ACM SIGCAS Computers and Society*. 2014; 44 (2): 16-19.
7. Milgram S. Behavioral study of obedience. *Journal of Abnormal and Social Psychology*. 1963; 67 (4): 371–378. doi:10.1037/h0040525. PMID 14049516.
8. Wendorf C. History of American morality research, 1894–1932. *History of Psychology*. 2001; 4 (3): 272–288.
9. Field H. A Priority as an evaluative notion. In: Boghossian P, Peacocke C, editors. *New essays on the a priori*. New York, NY: Oxford University Press; 2000. pp. 177-148.
10. Haidt J, Rosenberg E, Hom H. Differentiating diversities: Moral diversity is not like other kinds. *Journal of Applied Social Psychology*. 2003; 33 (1): 1–36.
11. Iyengar S, Westwood SJ. Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*. 2015; 59(3): 690-707.
12. Hatemi PK, Funk CL, Medland SE, Maes HM, Silberg JL, Martin NG, et al. Genetic and environmental transmission of political attitudes over a life time. *The Journal of Politics*. 2009; 71(3): 1141-1156.
13. Cardinale, BJ, Emmett Duffy J, Gonzalez A, Hooper DU, Perrings C, Venail P, et al. Biodiversity loss and its impact on humanity. *Nature*. 2012; 486: 59–67. doi:10.1038/nature11148.
14. Wilde DJ. Using student preferences to guide design team composition. In: *Proceedings of DETC '97*. New York, NY: ASME; 1997.
15. Wilde DJ. *Teamology: The Construction and Organization of Effective Teams*. New York, NY: Springer; 2008.
16. Walker L, Frimer J, Dunlop W. Varieties of moral personality: beyond the banality of heroism. *Journal of Personality*. 2010; 78 (3): 907–942. doi:10.1111/j.1467-6494.2010.00637.x. PMID 20573130.
17. Milgram S. Behavioral study of obedience. *Journal of Abnormal and Social Psychology*. 1963; 67 (4): 371–378. doi:10.1037/h0040525. PMID 14049516.
18. Sefton M, Shupp R, Walker J. The effect of rewards and sanctions in provision of public goods. *Economic Inquiry*. 2007; 45 (4): 671–690.
19. Aquino K, Reed A. The self-importance of moral identity. Journal of *Personality and Social Psychology*. 2002; 83 (6): 1423-1440.
20. Davis M. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*. 1983; 44 (1): 113-126.
21. Graham J, Haidt J, Nosek B. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*. 2009; 96 (5): 1029–1046. doi:10.1037/a0015141. PMID 19379034.
22. Levenson M, Kiehl K, Fitzpatrick C. Assessing psychopathic attributes in a noninstitutionalized population. *Journal of Personality and Social Psychology*. 1995; 68 (1): 151-158.
23. Lind G. Wie misst man moralisches urteil? Probleme und alternative möglichkeiten der messung eines komplexen konstrukts. In: Portele G, editor. *Sozialisation und Moral*. Weinheim: Beltz; 1978. pp. 171-201.
24. Rest J. *Development in Judging Moral Issues*. Minneapolis, MN: University of Minnesota Press; 1979.
25. Maydeu-Olivares A, Brown A. Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research*. 2010; 45 (6): 935-974.
26. Berinsky A, Huber G, Lenz G. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*. 2012; 20 (3): 351-368.
27. Buhrmester M, Kwang T, Gosling S. Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*. 2011; 6 (1): 3-5.
28. Paolacci G, Chandler J, Ipeirotis P. Running experiments on Amazon Mechanical Turk. *Judgment and Decision making*. 2010; 5 (5): 411-419.
29. Graham J, Haidt J, Nosek B. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*. 2009; 96 (5): 1029–1046. doi:10.1037/a0015141. PMID 19379034.
30. Rammstedt B, John O. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*. 2007; 41: 203–212.
31. Lind G. Wie misst man moralisches urteil? Probleme und alternative möglichkeiten der messung eines komplexen konstrukts. In: Portele G, editor. *Sozialisation und Moral*. Weinheim: Beltz; 1978. pp. 171-201.
32. Kohlberg L. The claim to moral adequacy of a highest stage of moral judgment. *The journal of philosophy*. 1973: 630-646.

33. Rest J. *Development in Judging Moral Issues*. Minneapolis, MN: University of Minnesota Press; 1979.
34. Rest J, Narvaez D, Thoma S, Bebeau M. DIT2: Devising and testing a revised instrument of moral judgment. *Journal of Educational Psychology*. 1999; 91 (4): 644-659.
35. Berinsky A, Huber G, Lenz G. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*. 2012; 20 (3): 351-368.
36. Lind G. The cross-cultural validity of the Moral Judgment Test: Findings from 27 cross-cultural studies. In: Presentation at the conference of the American Psychological Association; 2005. pp. 18-21.
37. Graham J, Haidt J, Nosek B. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*. 2009; 96 (5): 1029–1046. doi:10.1037/a0015141. PMID 19379034.
38. Graham J, Nosek B, Haidt J, Iyer R, Koleva S, Ditto P. Mapping the moral domain. *Journal of Personality and Social Psychology*. 2011; 101 (2): 366-85. doi: 10.1037/a0021847
39. Holland J. Making Vocational Choices: A Theory of Vocational Personalities and Work Environments. Lutz, FL: *Psychological Assessment Resources;* 1997.
40. Santos-Lang C. Moral ecology approaches to machine ethics. In: van Rysewyk S, Pontier M, editors. *Machine Medical Ethics*. New York: Springer; 2014. pp. 111-127.
41. Wilde DJ. Using student preferences to guide design team composition. In: *Proceedings of DETC '97*. New York, NY: ASME; 1997.
42. Wilde DJ. *Teamology: The Construction and Organization of Effective Teams*. New York, NY: Springer; 2008.
43. Walker L, Frimer J, Dunlop W. Varieties of moral personality: beyond the banality of heroism. *Journal of Personality*. 2010; 78 (3): 907–942. doi:10.1111/j.1467-6494.2010.00637.x. PMID 20573130.
44. Milgram S. Behavioral study of obedience. *Journal of Abnormal and Social Psychology*. 1963; 67 (4): 371–378. doi:10.1037/h0040525. PMID 14049516.
45. Sefton M, Shupp R, Walker J. The effect of rewards and sanctions in provision of public goods. *Economic Inquiry*. 2007; 45 (4): 671–690.
46. Iyengar S, Westwood SJ. Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*. 2015; 59(3): 690-707.

Appendix A: The GRINSQ

The GRINSQ consists of the following 24 questions. Some are very similar to others—that is intentional. For each statement, circle the letter (A or B) for the ending with which you agree the most:

1. My higher priority in life is…….….	A. ...to know what I am trying to achieve. B. ...to discover new possibilities.
2. In the future, I will.....	A. ...focus on the people I love the most. B. ...stay pure.
3. My higher priority in life is…….….	A. ...to know what I am trying to achieve. B. ...to serve something greater than myself.
4. I cannot be my best self.....	A. ...when my work does not require creativity. B. ...when I cannot empathize.
5. My higher priority in life is……..	A. ...to get results. B. ...to be lovable.
6. People who have known me longest treat me as if I am more……..	A. ...morally strict. B. ...idealistic but impractical.
7. I cannot be my best self…….…….	A. ...when my work does not require creativity. B. ...when I do not know the criteria by which success is measured.
8. My higher priority in life is……..	A. ...to serve something greater than myself. B. ...the feelings of the people closest to me.
9. I am more concerned about stress which...	A. ...leads me to experiment with less-pure behaviors. B. ...puts my plans on hold.
10. My higher priority in life is…….	A. ...the feelings of the people closest to me. B. ...to discover new possibilities.
11. In the future, I will…….....	A. ...focus on the people I love the most. B. ...make measurable achievements.
12. My higher priority in life is……..	A. ...to discover new possibilities. B. ...to serve something greater than myself.
13. I would prefer to have…….…….	A. ...unquestioned authorities. B. ...no plan.
14. My higher priority in life is……..	A. ...to be lovable. B. ...to exercise self-discipline.
15. In the future, I will……	A. ...make measurable achievements. B. ...stay pure.
16. More than others do, I……..	A. ...question existing best practices. B. ...maintain relationships.
17. My higher priority in life is……..	A. ...to know what I am trying to achieve. B. ...the feelings of the people closest to me.
18. People are more likely to complain about……..	A. ...my old-fashioned morals. B. ...my far-fetched proposals.
19. I am more concerned about stress which...	A. ...blocks my creativity. B. ...puts my plans on hold.
20. More than others do, I……...	A. ...uphold moral principles. B. ...maintain relationships.
21. My higher priority in life is……..	A. ...to exercise self-discipline. B. ...to get results.
22. In the future, I will……	A. ...focus on the people I love the most. B. ...convince others to open their minds.
23. I would prefer to have……...	A. ...no plan. B. ...a "pure business" culture.
24. More than others do, I……...	A. ...question existing best practices. B. ...uphold moral principles.

Scoring: G = (1B + 4A + 6B) + (7A + 10B + 12A) + (13B + 16A + 18B) + (19A + 22B + 24B); R = (2A + 4B + 5B) + (8B + 10A + 11A) + (14A + 16B + 17B) + (20B + 22A + 23A); I = (2B + 3B + 6A) + (8A + 9A + 12B) + (14B + 15B + 18A) + (20A + 21A + 24B); N = (1A + 3A + 5A) + (9B + 7B + 11B) + (13A + 15A + 17A) + (19B + 21B + 23B)

Appendix B: GRIN frequencies in different subsamples

		Gadfly > 8		*Rel. > 8*		*Inst. > 8*		*Neg. > 8*		*Not Identifiable*	
Sample	*N*	*%*	*φ*	*%*	*φ*	*%*	*φ*	*%*	*φ*	*%*	*φ*
Total sample	250	18%		30%		9%		26%		27%	
Male	120	17%	-0.02	20%	-0.22**	5%	-0.13*	29%	0.08	34%	0.15*
Heterosexual	217	15%	-0.16*	29%	-0.08	10%	0.08	26%	0.04	29%	0.08
Age											
18-24	58	17%	-0.01	21%	-0.12	7%	-0.04	29%	0.05	33%	0.07
25-34	94	19%	0.03	31%	0.01	7%	-0.04	26%	0.00	28%	0.01
35-49	50	16%	-0.02	36%	0.06	4%	-0.08	28%	0.03	26%	-0.01
50+	48	17%	-0.01	35%	0.05	19%	0.17**	19%	-0.08	21%	-0.07
Political orientation											
Conservative	49	6%	-0.15*	24%	-0.06	29%	0.34**	24%	-0.01	22%	-0.05
Moderate	59	15%	-0.03	27%	-0.04	7%	-0.04	25%	0.00	32%	0.06
Liberal	142	23%	0.15*	34%	0.08	3%	-0.24**	26%	0.01	27%	0.01
Civics/politics is part of my identity	49	24%	0.09	20%	-0.11	6%	-0.05	41%	0.17**	24%	-0.03
Holland code											
Artistic	15	47%	0.19**	27%	-0.02	0%	-0.08	20%	-0.03	20%	-0.04
Conventional	57	18%	0.00	33%	0.03	12%	0.07	35%	0.12	18%	-0.12
Enterprising	44	14%	-0.05	34%	0.04	11%	0.04	39%	0.14*	16%	-0.12
Investigative	40	25%	0.08	33%	0.02	8%	-0.02	28%	0.02	20%	-0.07
Realistic	43	21%	0.04	35%	0.04	9%	0.01	16%	-0.10	23%	-0.04
Social	45	18%	0.00	36%	0.05	7%	-0.04	22%	-0.04	27%	0.01
Religion											
Christian	107	5%	-0.29**	37%	0.13*	16%	0.25**	22%	-0.06	26%	-0.02
Converted Christian	33	0%	-0.18**	30%	0.00	39%	0.42**	6%	-0.17**	30%	0.03
No religion	104	29%	0.25**	26%	-0.08	3%	-0.18**	31%	0.10	12%	-0.01
Converted non-Christian	31	35%	0.18**	26%	-0.04	6%	-0.03	32%	0.06	13%	-0.12
Family											
Romance is part of my identity	74	16%	-0.02	45%	0.20**	8%	-0.02	24%	0-.02	22%	-0.08
Responsible to raise children	84	20%	0.05	38%	0.12	7%	-0.04	26%	0.01	21%	-0.09
Child care is part of my identity	60	18%	0.01	48%	0.22**	7%	-0.04	25%	-0.01	17%	-0.13*
Still dependent after age 25	30	23%	0.06	37%	0.05	7%	-0/03	10%	-0.13*	30%	0.02
Team sport is part of my identity	29	10%	-0.07	10%	-0.16*	3%	-0.07	34%	0.07	41%	0.12
Accused of a crime or other serious betrayal of trust	41	29%	0.14*	15%	-0.15*	2%	-0.10	27%	0.01	39%	0.12

* $p<0.05$, ** $p<0.01$

Appendix C: History of Peer-Review

4/20/2014	Submitted to *Personality and Individual Differences*
5/20/2014	Rejected from *Personality and Individual Differences* with two reviewers. **JUSTIFICATION:** Recommending longer explanation. No suggested changes to method, but longer version does not fit journal's word limit.
5/26/2014	Submitted to *Journal of Personality and Social Psychology: Personality Processes and Individual Differences*
6/5/2014	Rejected by *Journal of Personality and Social Psychology* without review. **JUSTIFICATION:** Journal does not publish articles with only a single study, a new scale, and no behavioral measures
7/5/2014	Submitted to *Psychological Assessment*
7/8/2014	Rejected by *Psychological Assessment* without review. **JUSTIFICATION:** Not in the scope of the journal
7/10/2014	Submitted to *Journal of Personality Assessment*
7/13/2014	Rejected from *Journal of Personality Assessment* without review. **JUSTIFICATION:** Not in the scope of the journal
7/13/2014	Submitted to *Assessment*
9/29/2014	Rejected from *Assessement*. **JUSTIFICATION:** Unable to find more than one reviewer, and reviewer judges hypothesis unworthy of testing. No suggested changes to method
1/23/15	Submitted to *Cognitive Psychology*
1/26/15	Rejected by *Cognitive Psychology* without review. **JUSTIFICATION:** Makes "no substantial contribution to theory"
2/3/2015	Submitted to *Journal of Research in Personality*
2/10/2015	Rejected by *Journal of Research in Personality* without review. **JUSTIFICATION:** Journal does not publish articles which introduce new scales and lack behavioral measures.
2/21/2015	Submitted to the *Journal of Cognition and Culture*
10/19/15	Rejected by the *Journal of Cognition and Culture*. **JUSTIFICATION:** Unable to find more than one reviewer, and reviewer believes survey methods are unreliable in general. No suggested changes to method
11/15/2015	Submitted to *PLOS ONE*
12/23/2015	Rejected by *PLOS ONE*. **JUSTIFICATION:** No reviewer could be found
12/28/2015	Appeal submitted to *PLOS ONE*
1/19/2016	Submitted to Peerage of Science
1/20/2016	*PLOS ONE* agrees to find reviewers
1/28/2016	*PLOS ONE* editor requests modifications
2/17/2016	No reviewers found on Peerage of Science
2/28/2016	Submitted to *PLOS ONE*
4/11/2016	Rejected by *PLOS ONE*. **JUSTIFICATION:** Quoted below. No suggested changes to method.
4/18/2016	Appeal submitted to *PLOS ONE*
9/29/2016	Appeal rejected by *PLOS ONE*. **JUSTIFICATION**: Quoted below. No suggested changes to method.

The first four blind reviews are not included below, since they were for earlier versions. What follows are the blind reviews and appeal for publication in PLOS ONE. In the end, the senior editor rejected the appeal, but did not address any of the arguments made in the appeal nor specify any changes that should be made to the method for testing the hypothesis.

4/11/16 From PLOS ONE

Dear Mr. Santos-Lang,

Thank you for submitting your manuscript to PLOS ONE. After careful consideration, we have decided that your manuscript does not meet our criteria for publication and must therefore be rejected.

Specifically:
Both reviewers and myself consider that your article presents numerous statements that are not supported by the data available, in your studies but also in the rest of the literature.

I am sorry that we cannot be more positive on this occasion, but hope that you appreciate the reasons for this decision.

Yours sincerely,

Academic Editor
PLOS ONE

Comments to the Author

1. Is the manuscript technically sound, and do the data support the conclusions?

The manuscript must describe a technically sound piece of scientific research with data that supports the conclusions. Experiments must have been conducted rigorously, with appropriate controls, replication, and sample sizes. The conclusions must be drawn appropriately based on the data presented.

Reviewer #1: No
Reviewer #2: Partly

2. Has the statistical analysis been performed appropriately and rigorously?

Reviewer #1: I Don't Know
Reviewer #2: No

3. Does the manuscript adhere to the PLOS Data Policy?

Authors must follow the PLOS Data policy, which requires authors to make all data underlying the findings described in their manuscript fully available without restriction. Please refer to the author's Data Availability Statement in the manuscript. All data and related metadata must be deposited in an appropriate public repository, unless already provided as part of the submitted article or supporting information. If there are restrictions on the ability of authors to publicly share data—e.g. privacy or use of data from a third party—these reasons must be specified.

Reviewer #1: Yes

Reviewer #2: Yes

4. Is the manuscript presented in an intelligible fashion and written in standard English?

PLOS ONE does not copyedit accepted manuscripts, so the language in submitted articles must be clear, correct, and unambiguous. Any typographical or grammatical errors should be corrected at revision, so please note any specific errors here.

Reviewer #1: No
Reviewer #2: No

5. Review Comments to the Author

Please use the space provided to explain your answers to the questions above. You may also include additional comments for the author, including concerns about dual publication, research ethics, or publication ethics. (Please upload your review as an attachment if it exceeds 20,000 characters)

Reviewer #1:

This manuscript reports an instrument, the Gadfly-Relational-Institutional-Negotiator Self-Quiz, which is purported to measure “typologies that apply universally, much as the periodic table of elements applies to atoms on all planets” (p. 2). The author first applied the typologies of gadfly, relational, institutional, and negotiator to computer modules, and in this paper he applies them to humans. An inventory based on these four types is generated and evidence intended to validate it is supplied.

I do not think that this manuscript is suitable for publication for reasons that are specified in this paragraph and the next, and in several of my comments. I cannot see how someone can make claims of “universal interdependent types” without having evidence from the universe. It seems to me that one’s arguments would have to be restricted to societies of the world. The manuscript gets off to a horrible start with the first two sentences of the abstract: “Science fiction about intelligent aliens has long imagined a science of sociology with typologies that apply universally, much as the periodic table of the elements applies to atoms on all planets. The GRIN model purports to offer such a universal typology.” This statement strikes me as outrageous, and it will immediately trigger to readers that the author should not be taken seriously. Why are we not considering Abnegation (the selfless), Amity (the peaceful), Candor (the honest), Dauntless (the brave), and Erudite (the intellectual), the five types from the Divergent science fiction movie series? My guess is that I could construct an instrument for humans that would classify people into these five types just as readily as the current instrument classifies them into the four GRIN categories. Is the GRIN classification any more valid than the Divergent classification, and does it provide any more insight into human behavior?

A more appropriate way to set the stage for the paper would be to lay out in detail the evidence and arguments for the four types in machines and then develop out of it a case for why they should extend to humans. There would then need to be detailed consideration as to how the GRIN classification compares to instruments that categorize humans in other ways, rather than the somewhat superficial comparison that is done in the current manuscript. I find it a little troubling when the statement is made, “Most human brains have regulatory mechanisms… which could prevent manifestation of a GRIN-type in a given context, thus forcing a human to (temporarily) manifest a different

type." If the GRIN types are context-specific, of how much value will they be in predicting human behavior?

Comments:

1. I have never heard of the term "Aspies" before, but I surmised that it referred to those with Asperger's syndrome. A check of Wikipedia confirmed this. The full term should be used in a paper like this instead of a slang name that could be taken to be derogatory.

2. Evaluativism, par. 1, line 7 – "adaption" should be "adaptation"; in general, "evaluatism is a term with which I am not familiar. Although I can get the general idea from the paragraph, it needs to be described in more detail.

3. Evaluatism, par. 3 – GRIN-type predicts evaluatism, but what would be important is if other more plausible schemes would not also predict it.

4. Current study, par. 1, line 8 – What "the required distributions" are is not clear. Is this referring to the bimodal distributions mentioned in the third line?

5. Selection of Assessment Type – This section cites one reference on Likert-type and forced-choice questions, but the basis for the criticisms of Likert-type scales is not clear.

6. Materials and Procedure, pars. 3 and 4 – It is not clear why machine types were the focus of the examples provided by the SME's and the content assumptions.

7. Content Validity, par. 2 – These values of mean SME agreement do not seem very high to me, as the highest value of 3.9 does not attain the weakest of the two "agree" ratings (4 and 5).

8. Content Validity, the four bullets – The four types here are identified and discussed with respect to machines rather than humans. It is stated after the last part of Table 2 that "six of the nine SMEs expressed serious caution about the assumptions required to generate these ratings, asserting that personhood goes beyond mechanical phenomena." This warning is then dismissed as stemming from "threat to human egos". The warning needs to be taken much more seriously, and more formal statistical and logical analyses need to be reported for such dismissal to be warranted.

Reviewer #2:

The manuscript titled "Measuring Evaluative Computational Differences in Humans" is an interesting study designed to validate the properties of a scale that contains four dimensions: Gadfly, relational, institutional, and negotiator. The author also presents some results about how his scale is related to other measures including moral judgment, moral foundations, and personality traits. Unfortunately, I do have reservations and I would recommend that this study not be published in PLOS ONE.

To begin, there literature review fails to provide the foundation necessary to understand or evaluate the outcomes in this study. Typically with a scale development study, the literature review is designed to identify a theoretical or conceptual position (in this case, a justification is required for the four dimensions of gadfly, relational, institutional, and negotiator) so that the reader can understand why these dimensional can be used to measure a key outcome or construct. The literature review fails to either justify or review these dimensions clearly. I also had difficulty mapping the references in the literature review onto the purpose of the

study.

Next, the methods describe how the theoretical or conceptual position is to be evaluated. The author does provide a good summary of his sample and his procedures. But he does not provide a complete description of his method or his results. For example, in the content-related validity section the content assumptions and rationales list how the items align with the construct. But there is no description of why the rationales are included or where these rationales came from. A detailed justification for these rationales is critical for understanding the item development and evaluation approach used on the GRINSQ scale. Also, the reliability estimates are presented for the GRINSQ, but also for the MFQ, BFI-10, and MJT. First, the reliability for the established scales (i.e., the later three) should be presented from a norming sample, not based on his convenience sample. Second, the reliability results are averaged. I don't understand what an average reliability represents. Reliability is a composite measure of the item-level results. It does not require an average. In the structural validity section, the author presents exploratory factor analytic results without describing the extraction or rotation method. However, the more appropriate analysis would be a confirmatory factor analysis given the GRINSQ scales and structures have already been defined. Finally, the author presents pragmatic validity findings. Pragmatic validity is not a type of validity. I think the author is describing either predictive or criterion-related validity i.e., how the GRINSQ correlates with other psychological measures. Regardless, this section must be guided by a detailed description of how and why different types of associations and relationships would occur. In this manuscript, the correlations are largely exploratory and interpreted post hoc without any clear justifications.

To summarize, scale development and validation is an important methodological activity. The author does provide evidence to validate his scale. However, a logical structure is needed to align the scale concepts with the analysis and the results so that the outcomes of the validation task can be interpreted. Unfortunately, this study lacks the structure to clearly interpret the outcomes from the analyses.

4/17/16 To PLOS

I would like to appeal the decision not to publish my submission to PLOS ONE, PONE-D-16-06252: "Measuring Evaluative Computational Differences in Humans."

I deeply appreciate the work of the Editor and the reviewers. Their comments demonstrate that the reviewers are both thoughtful and knowledgeable, so it is encouraging to see that they were unable to identify any legitimate reason to reject the article. However, it is possible that I am misunderstanding one or more of their comments, and I hope you will tell me if that is the case. I have copied their review below (in bold) with my response to each claim below it:

Reviewer #1: This manuscript reports an instrument, the Gadfly-Relational-Institutional-Negotiator Self-Quiz, which is purported to measure "typologies that apply universally, much as the periodic table of elements applies to atoms on all planets" (p. 2). The author first applied the typologies of gadfly, relational, institutional, and negotiator to computer modules, and in this paper he applies them to humans. An inventory based on these four types is generated and evidence intended to validate it is supplied.

I do not think that this manuscript is suitable for publication for reasons that are specified in this paragraph and the next, and in several of my comments. I cannot see how someone

can make claims of "universal interdependent types" without having evidence from the universe. It seems to me that one's arguments would have to be restricted to societies of the world.

[A] I take the reviewer's point to be that one could never prove a universal claim based purely on evidence from Earth. This reason is illegitimate in two ways: First, although data gathered purely on Earth may never *prove* any universal hypothesis, it can nonetheless *support* such hypotheses, and that is all that the official publication criteria require (which is why PLOS ONE can publish experiments relevant to universal claims in chemistry and physics). Second, the conclusion of the article (quoted as follows) is very careful to distinguish the hypothesis of universality from the hypothesis that GRIN-types exist among humans, and to acknowledge the different relationships the data has to each hypothesis:

> The primary research question for this study was whether it is possible to discern GRIN-types in humans. We were able to develop a reliable self-quiz, the GRINSQ, which has the expected structure, scope of content, and relationships with the BFI-10 and MFQ. This supports the conclusion that GRIN-types exist among humans, and is a step towards establishing GRIN-type as a universal social typology.

The manuscript gets off to a horrible start with the first two sentences of the abstract: "Science fiction about intelligent aliens has long imagined a science of sociology with typologies that apply universally, much as the periodic table of the elements applies to atoms on all planets. The GRIN model purports to offer such a universal typology." This statement strikes me as outrageous, and it will immediately trigger to readers that the author should not be taken seriously.

[B] Rather than specify something outrageous beyond [A], the reviewer here offers the distinct argument that the article should be rejected because people will not take it seriously. This is merely one person's opinion, and not one of the publication criteria. PLOS ONE's publication policy is to publish all rigorous science online and let the readers decide for themselves which articles they consider noteworthy.

Why are we not considering Abnegation (the selfless), Amity (the peaceful), Candor (the honest), Dauntless (the brave), and Erudite (the intellectual), the five types from the Divergent science fiction movie series? My guess is that I could construct an instrument for humans that would classify people into these five types just as readily as the current instrument classifies them into the four GRIN categories. Is the GRIN classification any more valid than the Divergent classification, and does it provide any more insight into human behavior?

[C] Whether the reviewer can create a valid Divergent classification has no bearing, of course, on whether the current study should be published. Here the reviewer seems to be offering only the observation that the validity of any experiment (including this one) does not guarantee its noteworthiness. PLOS ONE's publication policy is to publish all rigorous science online and let the readers decide for themselves which articles they consider noteworthy.

A more appropriate way to set the stage for the paper would be to lay out in detail the evidence and arguments for the four types in machines and then develop out of it a case for why they should extend to humans. There would then need to be detailed consideration as to how the GRIN classification compares to instruments that categorize humans in other ways, rather than the somewhat superficial comparison that is done in the current manuscript.

[D] The reviewer appears to be asking for reorganization or additional detail not required to reproduce the experiment and therefore not required for publication. I appreciate specific suggestions about how best to tell a story (the reviewer offers some below), but style is subjective. The objective criterion for publication is whether the article makes accessible a valid

reproducible experiment. If that is achieved, the story can be told in other publications in a variety of styles--there is no need to please every sense of style in this one article.

I find it a little troubling when the statement is made, "Most human brains have regulatory mechanisms… which could prevent manifestation of a GRIN-type in a given context, thus forcing a human to (temporarily) manifest a different type." If the GRIN types are context-specific, of how much value will they be in predicting human behavior?
[E] I think the reviewer is raising the concern that the article might not be noteworthy, since context may cause people to behave counter to type. The extent to which this occurs remains to be seen. PLOS ONE's publication policy is to publish all rigorous science online and let the readers decide for themselves which articles they consider noteworthy.

Comments:

1. I have never heard of the term "Aspies" before, but I surmised that it referred to those with Asperger's syndrome. A check of Wikipedia confirmed this. The full term should be used in a paper like this instead of a slang name that could be taken to be derogatory.
[F] It is my understanding that some Aspies are offended by the term "those with Asperger's syndrome" since they successfully fought to remove "Asperger's syndrome" from the Diagnostic Manual. The reviewer is not providing reason for rejection here; it seems to be merely a helpful suggestion.

2. Evaluativism, par. 1, line 7 – "adaption" should be "adaptation";
[G] Technically, "adaption" and "adaptation" are synonyms, so either word will do. The reviewer is not providing reason for rejection here; it seems to be merely a helpful suggestion.

in general, "evaluatism is a term with which I am not familiar. Although I can get the general idea from the paragraph, it needs to be described in more detail.
[H] As in [D] above, the reviewer is seeking detail which is not required to replicate the experiment--there are other publications in which the curious can learn more about evaluativism.

3. Evaluatism, par. 3 – GRIN-type predicts evaluatism, but what would be important is if other more plausible schemes would not also predict it.
[I] The reviewer is referring to this passage:

> In addition to predicting the existence of evaluativism, the hypothesis that humans already specialize by GRIN-type predicts that evaluativism would handicap societies much as speciesism can handicap ecosystems (i.e. by leading to the disabling of components upon which the system as a whole depends [13]).

This is just background explaining the implications of the GRIN model, so the reviewer is not questioning the rigor of the actual experiment (which does not measure evaluativism). Neither is the reviewer claiming that the article makes any false claim. The reviewer is simply reminding us that this background does not prove the existence of GRIN types. The reviewer is merely offering an observation (which actually helps to motivate the experiment), not providing reason to reject the article.

4. Current study, par. 1, line 8 – What "the required distributions" are is not clear. Is this referring to the bimodal distributions mentioned in the third line?
[J] The reviewer is correct that "the required distributions" refers to bimodal distributions. The reviewer is not claiming that details required to replicate the experiment are unclear nor that the writing is unintelligible. This appears to be simply a helpful writing style suggestion for a background section, not a reason to reject.

5. Selection of Assessment Type – This section cites one reference on Likert-type and forced-choice questions, but the basis for the criticisms of Likert-type scales is not clear.
[K] As in [J] above, the reviewer seems to be offering a friendly writing suggestion by identifying a background section where he/she would like more detail. These are not details required to replicate the experiment, nor is the reviewer suggesting that this section is unintelligible. The section is as follows:

> Likert-type questions also have some important disadvantages [compared to forced-choice]. The first is that they support an illusion that one's construct is complete. A classifier based on forced-choice questions (e.g., a species classifier) leaves some subjects unidentified, thus exposing the incompleteness of the construct, and leaving a path to improve it.
>
> The second major disadvantage of Likert-type scales is their inability to distinguish subjects with more reliable results. If subjects hide their types (which we would expect, based on the sociological predictions), they might not do so equally. Forced-choice batteries allow us to assess reliability on a subject-by-subject basis, naturally classifying subjects with unreliable results as unidentified. Likert-type batteries, in contrast, misrepresent random answers as valid (i.e. falsely indicating balance).

6. Materials and Procedure, pars. 3 and 4 – It is not clear why machine types were the focus of the examples provided by the SME's and the content assumptions.
[L] As in [J] above, the reviewer seems to be offering a friendly writing suggestion by identifying a section in which he/she would like more detail. The "why" is not required to replicate the experiment, nor is the reviewer suggesting that this section is unintelligible. Furthermore, the "why" was answered when the GRIN-types construct was introduced, explaining that it was defined in its original publication in terms of machine types:

> ...computer modules were first sorted into these types in the field of machine ethics, where they were called "GRIN-types" [1]:
>
> Gadfly: Unpredictable due to use of novelty (e.g. an individual mutator in evolutionary computation) – compared to pragmatic ethics
> Relational: Unpredictable due to network effects (e.g. a cell of level 3 or 4 cellular automata) – compared to virtue ethics
> Institutional: Predictably upholds rules (e.g. a standard calculator) – compared to deontological ethics
> Negotiator: Predictably converges on maximizing a measurable goal (i.e. supervised machine learning for financial-trading) – compared to consequentialist ethics

7. Content Validity, par. 2 – These values of mean SME agreement do not seem very high to me, as the highest value of 3.9 does not attain the weakest of the two "agree" ratings (4 and 5).
[M] Although the reviewer seems tentative, the argument being implied here is that content validity cannot be established without high mean Subject Matter Expert (SME) agreement. The test of content validity is not a popularity contest. The article describes the logic of the content validity test:

> If the GRINSQ measured something other than proclivities for GRIN algorithm types, we would expect SMEs to agree about which items violate that intent. They had no such agreement.

One would like to have endorsements from SMEs, but passing the content validity test does not require that--it requires only that no mean is low.

8. Content Validity, the four bullets – The four types here are identified and discussed with respect to machines rather than humans. It is stated after the last part of Table 2 that "six of the nine SMEs expressed serious caution about the assumptions required to generate these ratings, asserting that personhood goes beyond mechanical phenomena." This warning is then dismissed as stemming from "threat to human egos". The warning needs to be taken much more seriously, and more formal statistical and logical analyses need to be reported for such dismissal to be warranted.
[N] The reviewer is referring to this passage:

> Via open-ended comment, six of the nine SMEs expressed serious caution about the assumptions required to generate these ratings, asserting that personhood goes beyond mechanical phenomena. This warning could stem from threat to human egos. Consistent with this hypothesis, SMEs agreed least with the comparison most threatening to human egos (i.e. that between humans and standard calculators).

The assertion that personhood goes beyond mechanical phenomena is controversial--maybe even mystical. It is an interesting controversy, but the criteria for publication of this experiment certainly do not require resolving that controversy. As stated in [M], popularity is not one of the tests of validity, so the endorsement of the SMEs is not relevant to whether the data support the conclusion of the article. The hypothesis that the assertion stemmed from threat to human egos was not presented as a conclusion of this article, so it need not include data to support it.

Reviewer #2: The manuscript titled "Measuring Evaluative Computational Differences in Humans" is an interesting study designed to validate the properties of a scale that contains four dimensions: Gadfly, relational, institutional, and negotiator. The author also presents some results about how his scale is related to other measures including moral judgment, moral foundations, and personality traits. Unfortunately, I do have reservations and I would recommend that this study not be published in PLOS ONE.

To begin, there literature review fails to provide the foundation necessary to understand or evaluate the outcomes in this study. Typically with a scale development study, the literature review is designed to identify a theoretical or conceptual position (in this case, a justification is required for the four dimensions of gadfly, relational, institutional, and negotiator) so that the reader can understand why these dimensional can be used to measure a key outcome or construct. The literature review fails to either justify or review these dimensions clearly. I also had difficulty mapping the references in the literature review onto the purpose of the study.
[O] As in [D] above, the reviewer appears to be asking for additional detail of background material not required to reproduce the experiment and therefore not required for publication. The publication criteria of PLOS ONE are substantially different from those of typical journals which require demonstration of noteworthiness, so this article need not be "typical" to pass the criteria for publication. The validity of an experiment does not require theoretical justification--as happened in the case of personality theory, phenomena may be discovered empirically before being explained theoretically. Secondly, theoretical justification for the GRIN typology *was* offered in the introduction as follows:

> One reason to expect that interdependent specialization to be advantageous in a society comes from the observation that rate of adaptation is limited by at least four distinct factors:
> 1. Rate at which novel configurations are produced

2. Selection pressure privileging better configurations
3. Fidelity with which proven configurations are reproduced
4. Network localization

Unless the best approaches for promoting these factors all happen to be the same, the most quickly adapting society will be specialized such that each member promotes only a subset of these factors, yet collectively the members promote them all.

Next, the methods describe how the theoretical or conceptual position is to be evaluated. The author does provide a good summary of his sample and his procedures. But he does not provide a complete description of his method or his results.

For example, in the content-related validity section the content assumptions and rationales list how the items align with the construct. But there is no description of why the rationales are included or where these rationales came from. A detailed justification for these rationales is critical for understanding the item development and evaluation approach used on the GRINSQ scale.
[P] This is not a claim that further detail would be required to replicate the tests described in this study. Replication would require copying the same questions word-for-word (i.e. the same GRINSQ items and rationales), so this article does not need to explain how to develop additional items or how to come up with rationales. The details being sought here are background, as in [D] above. No average SME agreement score was low, so the data confirms that none of the rationales given in this study were inappropriate.

Also, the reliability estimates are presented for the GRINSQ, but also for the MFQ, BFI-10, and MJT. First, the reliability for the established scales (i.e., the later three) should be presented from a norming sample, not based on his convenience sample.
[Q] This is another objection about background, rather than about methods or results. "Established scales" refers to scales for which reliability statistics (and other validity criteria) were already measured in prior studies using separate samples. This article cites those prior studies. As in [D] above, the preference to quote statistics from the prior studies (rather than have curious readers access them in context) is a style issue which does not constitute a legitimate reason to reject. This study additionally reports the reliability measures for the convenience sample, but there is nothing wrong with providing that additional information. In fact, it is necessary because any replication would be expected to include the same analysis as confirmation that its convenience sample was not biased in a way that would make the comparison scales unreliable.

Second, the reliability results are averaged. I don't understand what an average reliability represents. Reliability is a composite measure of the item-level results. It does not require an average.
[R] The reviewer is referring to Table 3 which reports Chronbach's alpha for each factor and reports an average for each scale. The reviewer is suggesting merely that the data for the averages be omitted--not that anything need be added. The sentence beginning "A 0.69 average was measured for the GRINSQ..." would instead be phrased "Alphas of 0.72, 0.69, 0.74 and 0.62 were measured for the GRINSQ". Those numbers are reported in the table, so this amounts to a criticism of style as in [D] above (when numbers are so close to their average, and the average is so different from the averages on other scales, I find averages to be an efficient way to communicate).

In the structural validity section, the author presents exploratory factor analytic results without describing the extraction or rotation method. However, the more appropriate analysis would be a confirmatory factor analysis given the GRINSQ scales and structures have already been defined.

[S] This is Reviewer #2's only specific argument that the analysis does not rigorously support the conclusion with data, but the reviewer appears to be confusing this study with a different one. This experiment includes no factor analyses at all, neither confirmatory nor exploratory. Factor analysis would be used to validate Likert-type scales, and most scales use Likert-type scales, but the GRINSQ uses forced-choice questions instead. Statisticians generally accept that force-choice questions need to be analysed in a different way, and the background section explains this (and provides citation):

> There were several reasons to choose forced-choice questions over Likert-type scales. The major advantage of Likert-type questions is to produce scalar numbers permitting analysis via correlation, factor-analysis, and regression [25], but this advantage is illusory with categorical constructs like GRIN-type and species. No matter how scalar one's measure of "humanness," for example, it would be invalid to call one person "more" human than another or to perform regressions which suggest strategies to become "more" human. Analyzing GRIN-type in degrees would be just as absurd.

Since use of factor analysis (or any other form of regression) would be grounds to reject this study, it instead employed a probability model to test structural validity:

> The test of structural validity for the GRINSQ is in the bimodal distributions of its factors...A score greater than ten would be considered statistically significant by itself ($p<0.015$) because the probability of generating a set of scores with at least one greater than ten is only 1.5%. Our sample included 58 subjects (23%) with such scores: 8 gadfly, 30 relational, 6 institutional, and 14 negotiator. Collectively, they confirm that at least four distinct orientations can be discerned among Mechanical Turkers in the United States.

Finally, the author presents pragmatic validity findings. Pragmatic validity is not a type of validity. I think the author is describing either predictive or criterion-related validity i.e., how the GRINSQ correlates with other psychological measures. Regardless, this section must be guided by a detailed description of how and why different types of associations and relationships would occur. In this manuscript, the correlations are largely exploratory and interpreted post hoc without any clear justifications.
[T] How the GRINSQ correlates with other psychological measures is given in the "Convergent/Divergent Validity" section (to which the reviewer raised no objections). The Pragmatic Validity section is explained as follows:

> Graham et al. [38] argued that new scales merit attention only if they allow us to support important new conclusions. For example, they demonstrated the pragmatic validity of the MFQ by showing that it allows scientists to explain the intractability of political disagreement as stemming from moral differences. Likewise, the GRINSQ supports new explanations of various phenomena. These explanations are supported through chi-squared test as tabulated in S2 Appendix B.

It may be true that most validation studies do not include tests of pragmatic validity, but the GRINSQ is compared to the MFQ, and it can't hurt to subject it to all the same tests. It is appropriate that the statistical analysis in this section is exploratory and interpreted post hoc, since the test of pragmatic validity requires merely that the GRINSQ produces data for which new explanations become needed. If the reviewer considers this test unnecessary, he/she is welcome to ignore this section, so its inclusion is certainly not a legitimate reason to reject the article.

To summarize, scale development and validation is an important methodological activity. The author does provide evidence to validate his scale. However, a logical structure is needed to align the scale concepts with the analysis and the results so that the outcomes

of the validation task can be interpreted. Unfortunately, this study lacks the structure to clearly interpret the outcomes from the analyses.

I hope to acknowledge that the reviews were honest, intelligent, and thoughtful. Science is not easy, especially when crossing disciplines. It was the job of the reviewers to raise all objections that came to their minds, whether legitimate or not, and they fulfilled that responsibility admirably. Very few people would be able to come up with the objections these reviewers raised.

1. The seven publication criteria for PLOS ONE are:
2. The study presents the results of primary scientific research.
3. Results reported have not been published elsewhere.
4. Experiments, statistics, and other analyses are performed to a high technical standard and are described in sufficient detail.
5. Conclusions are presented in an appropriate fashion and are supported by the data.
6. The article is presented in an intelligible fashion and is written in standard English.
7. The research meets all applicable standards for the ethics of experimentation and research integrity.
8. The article adheres to appropriate reporting guidelines and community standards for data availability.

The reviewers raised no specific challenges to 1, 2, 3, 5, 6, or 7. The challenges to 4 were [A], [M], [N] and [S] which relied on the following falsehoods:

- [A] It is impossible to study universal claims empirically on Earth
- [M] & [N] The test of content validity is a popularity contest among experts
- [S] This study used exploratory factor analysis

Although both reviewers deny "5. The article is presented in an intelligible fashion and is written in standard English," neither identifies any specific violation of this criteria. The English language has rules, so one should be able to point to at least one specific violated rule if this criterion is not met, and it is reasonable for the author to expect that level of evidence. To say merely "paragraph X isn't clear to me" can be a helpful comment, but it doesn't necessarily mean the paragraph is unintelligible (e.g. one may simply be looking for additional detail), so it does not qualify as evidence that the article is unintelligible. PLOS ONE has objective publication criteria, but the reviewers may have reinterpreted #5 as the subjective criterion "Is the article is well written?" They did not agree on any particular section as unclear, nor name any rule of English as violated.

The other challenges were stylistic suggestions ([D], [F], [G], [H], [J], [K], [L], [O], [P], [Q], [R], and [T]), and complaints that the research might not be noteworthy even if rigorous ([B], [C], and [E]). If the author's goal were to be accepted by the scientific community, then these reviewers would be a focus-group, and the latter kinds of critiques would be important. However, this article represents an effort to reform certain scientific fields, and reform rarely comes from within. The reviewers are not part of the intended audience--they are part of the expected opposition. For the purposes of this article, peer-reviewers function less like a focus-group than like a firing squad testing prototype body armor.

Here is my request to you:

1. If you, or the reviewers, can identify problems with my critique of the reviews, then please tell me what those problems are (so I can be corrected)

2. If you think there are legitimate rejection reasons not specifically articulated by the reviewers, then please send it for re-review so other reviewers can articulate them--in this case you would be expecting the re-reviewers to reject the article
3. If you see no legitimate reasons to reject even after these reviewers have worked so hard to find one, then please publish the article, thus opening a debate in which any other reader can raise any additional issues they may see.

Typically, if you receive an appeal for which the reviews did not identify legitimate reason to reject, you probably assume that the reviewers failed to do their job. In this case, however, the better explanation for why the reviewers would reject without legitimate reason is that science has politics, and this article challenges the status quo. The status quo would lead reviewer #2 to expect factor-analysis to be used and cause reviewer #1 to write that the first sentence of the abstract "will immediately trigger to readers that the author should not be taken seriously." This article is so deeply irreverent that review might as well not be blind--the article obviously comes from an outsider. If you take bias seriously enough to practice blind review, yet blinding won't neutralize bias against this article, then you probably ought not expect positive reviews, even in re-review.

Here we have a conflict between the current scientific community and an outsider who seeks peaceful resolution of disagreement by constructively offering a new tool which scientists are free to test for themselves. How should PLOS ONE handle such a situation?

You could let the data decide. If thoughtful expert reviewers cannot give legitimate reason to reject an article, then the data best supports the conclusion that the submission meets the publication criteria. Had there been any specific challenge which was raised by multiple reviewers, one might focus on that particular criticism as a likely genuine problem, but the reviewers offered no such agreement. It really looks like any legitimate flaws in this study are non-obvious, so it deserves to be brought into open debate.

To refuse to publish in such situations could be considered dishonest: The publication criteria do not include "8. It is endorsed by peer-reviewers." The first time I submitted this article, it was rejected purely on the basis that no editor wanted to take it on, and I appealed on the basis that finding an editor is not one of the criteria. Likewise, getting endorsements from peer-reviewers is not part of the criteria. By advertising publication criteria and claiming to have an appeal process, PLOS ONE claims that any submission it rejects fails to meet one or more of the advertised criteria--you would knowingly be making a false claim in your advertisement, if you refuse to publish even though you believe the reviewers have no legitimate reason to reject.

On the other hand, if PLOS ONE handles such situations by publishing such controversial studies, and even one such study succeeds at bringing about a paradigm shift, then PLOS ONE will go down in history as a hero of the scientific process. The work of Dr. Courvoiser and the reviewers was heroic. They were honest, thoughtful, and dutifully took on a task which needed to be done. I hope the appeal process will match that heroism.

Thank you for your consideration.

Sincerely,

Chris Santos-Lang

9/13/16 From PLOS ONE

Dear Mr. Santos-Lang,

I am writing to you with regard to your appeal on the editorial decision for your submission to PLOS ONE above. I apologize for the delay following up on this matter.

After careful consideration of the manuscript, the original reviews and your appeal letter, we consider that the decision for rejection should be upheld.

As you are aware, manuscripts submitted to PLOS ONE are assessed mainly on the basis of the scientific soundness of the work. However, as we understand the reports, the reviewers and Academic Editor have raised overlapping concerns on the contents of the manuscript in particular regarding the validation of the new tool presented and whether the data support some of the conclusions made.

In the light of those concerns, we feel that the manuscript does not currently meet some of our criteria for publication including "Experiments, statistics, and other analyses are performed to a high technical standard and are described in sufficient detail", "Conclusions are presented in an appropriate fashion and are supported by the data" and specifically for manuscripts describing new tools "Submissions presenting methods, software, databases, or tools must demonstrate that the new tool achieves its intended purpose. If similar options already exist, the submitted manuscript must demonstrate that the new tool is an improvement over existing options".

I appreciate that you will be disappointed by this decision and I am sorry that we cannot be more positive on this occasion. Please note that decisions on appeal cases are final. I am sorry again for the delay in getting back to you and for the inconvenience caused.

Thank you for your interest in PLOS ONE.

Best wishes,

Senior Editor