The role of robotics and AI in technologically mediated human evolution: a constructive proposal	**Роль робототехники и ИИ в технологически опосредованной эволюции человека: конструктивное предложение**
Jeffrey White1,2	
Received: 30 January 2018 / Accepted: 26 December 2018	
Abstract	Аннотация
This paper proposes that existing computational modeling research programs may be combined into platforms for the information of public policy. The main idea is that computational models at select levels of organization may be integrated in natural terms describing biological cognition, thereby normalizing a platform for predictive simulations able to account for both human and environmental costs associated with different action plans and institutional arrangements over short and long time spans while minimizing computational requirements. Building from established research programs, the proposal aims to take advantage of current momentum in the direction of the integration of the cognitive with social and natural sciences, reduce start-up costs and increase speed of development. These are all important upshots given rising unease over the potential for AI and related technologies to shape the world going forward.	В этой статье предлагается, чтобы существующие исследовательские программы по вычислительному моделированию могли быть объединены в платформы для информирования государственной политики. Основная идея заключается в том, что вычислительные модели на отдельных уровнях организации могут быть интегрированы в естественных терминах, описывающих биологическое познание, тем самым нормализуя платформу для прогнозного моделирования, способную учитывать как человеческие, так и экологические издержки, связанные с различными планами действий и институциональными механизмами в течение короткого и длительного периода времени. время простирается при минимизации вычислительных требований. Основываясь на устоявшихся исследовательских программах, предложение направлено на то, чтобы воспользоваться нынешним импульсом в направлении интеграции когнитивных наук с социальными и естественными науками, снизить начальные затраты и увеличить скорость развития. Все это важные результаты, учитывая растущее беспокойство по поводу потенциала ИИ и связанных с ним технологий для формирования мира в будущем.
Keywords Cognitive social science · Computational model · Social simulation · Free energy principle · Directed evolution · AI arms race	Ключевые слова Когнитивные социальные науки · Вычислительная модель · Социальное моделирование · Принцип свободной энергии · Направленная эволюция · Гонка вооружений ИИ
1 Introduction	1 Введение
The potential for AI and related technologies to shape the future is increasingly an object of public and political concern. For instance, while addressing an audience of over 1 million on Knowledge Day, September 1, 2017, the President of the Russian Federation Vladimir Putin had this to say about AI and the future of humanity:	Потенциал искусственного интеллекта и связанных с ним технологий для формирования будущего становится все более предметом общественной и политической озабоченности. Например, обращаясь в День знаний, 1 сентября 2017 года, к аудитории, насчитывающей более 1 миллиона человек, Президент Российской Федерации Владимир Путин сказал следующее: об ИИ и будущем человечества:
Artificial intelligence is the future, not only for Russia, but for all humankind. It comes with colossal opportunities, but also threats that are difficult to predict. Whoever becomes the leader in this sphere will become	Искусственный интеллект - это будущее не только для России, но и для всего человечества. Это дает колоссальные возможности, но и угрозы, которые трудно предсказать. Кто бы ни стал лидером в этой сфере, он

the ruler of the world (as translated by Russia Today in RT 2017).	станет правителем мира (в переводе Russia Today в RT 2017).
Whoever leads AI will rule the world. In the current political context, the tendency for many readers in English may be to interpret such a statement as a threat, or at least as a horsewhip in an arms race pitting state actor against state actor in a zero-sum effort to optimize artificial intelligence in the industries of war (cf. Armstrong et al. 2016).	Тот, кто возглавит ИИ, будет править миром. В современном политическом контексте многие англоязычные читатели склонны интерпретировать подобное заявление как угрозу или, по крайней мере, как пришпоривание гонки вооружений, натравливающий одно государство на другое в стремлении с нулевой суммой оптимизировать искусственный интеллект в военной промышленности (ср. Армстронг и др. 2016).
However, there are other ways to see the future and the role of AI and robotics in shaping it (cf. White 2016). The purpose of this paper is to show that we need not anticipate an arms race eventuating in conflict, and rather that there is significant work ongoing in AI that points in the opposite direction. By recruiting and repurposing existing resources and established research programs, mutually beneficial ends may be identified, peaceful paths forward may be made explicit, and with these, rising anxieties due to currently resurgent geo-political polarization that might otherwise motivate a decision to initiate machine mediated conflict in resolution thereof may be quelled. Far from inviting mutual destruction, AI and related technologies may predicate open cooperation through thoroughly informed and mutually beneficial public policy, instead.	Однако есть и другие способы увидеть будущее и роль ИИ и робототехники в его формировании (см. Уайт 2016). Цель этой статьи - показать, что нам не нужно предвидеть гонку вооружений, которая может привести к конфликту, и что в ИИ ведется значительная работа, которая указывает в противоположном направлении. Путем рекрутирования и перепрофилирования существующих ресурсов и установленных исследовательских программ можно определить взаимовыгодные цели, четко обозначить мирные пути продвижения вперед, а вместе с этим можно подавить растущую тревогу, вызванную в настоящее время возрождающейся геополитической поляризацией, которая в противном случае могла бы мотивировать решение инициировать машинно-опосредованный конфликт для его разрешения. Вместо того, чтобы поощрять взаимное уничтожение, ИИ и связанные с ним технологии могут предвещать открытое сотрудничество посредством тщательно информированной и взаимовыгодной государственной политики.
The next section introduces Sun et al.'s innovative integration of the social with the cognitive sciences, beginning with his proposal that cognitive social models may be essential for understanding cognition, generally, and then turning to his suggestion that the study of such models should be part of the curriculum of policy studies given their unique potential to make explicit to policy makers the indirect consequences of policy changes. Section 3 briefly introduces theterms for translation between the cognitive, social and natural sciences required for the integration of models in a way that is informative to policy makers about policy impacts on individuals, communities and supporting ecologies, in the form of Friston et al.'s broad research program into the organizing principles of biological cognition. Section 4 briefly reviews three research programs in cognitive modeling, each at a selected level of organization complimenting the others in ways that, with results from	В следующем разделе представлена инновационная интеграция социальных с когнитивными науками Суна и соавторов, начиная с его предложения о том, что когнитивные социальные модели могут быть необходимы для понимания познания в целом, а затем обращаясь к его предположению, что изучение таких моделей должно быть частью учебной программы политических исследований, учитывая их уникальный потенциал, позволяющий прямо указывать политикам косвенные последствия политических изменений. В разделе 3 кратко представлены определения для перевода между когнитивными, социальными и естественными науками, необходимые для интеграции моделей таким образом, чтобы они были информативны для политиков о влиянии политики на отдельных лиц, сообщества и поддерживающие экологию, в форме Фристона и др.это обширная исследовательская

one level informing the next, may provide for the testing of policy changes over short and long time spans. Section 5 briefly sketches how these three programs may be integrated in informing transitions through critical periods, and the paper ends by indicating research required for such predictive simulations to inform public policy.	программа по изучению организующих принципов биологического познания. В разделе 4 кратко рассматриваются три исследовательские программы по когнитивному моделированию, каждая на выбранном уровне организации, дополняющая другие таким образом, что результаты одного уровня, информирующие следующий, могут обеспечить тестирование изменений политики в течение короткого и длительного промежутков времени. Раздел 5 кратко описывает, как эти три программы могут быть интегрированы в информирование переходов через критические периоды, а в конце документа указываются исследования, необходимые для такого прогнозного моделирования для информирования государственной политики.
2The call to integration	2 Призыв к интеграции
In redress of prior schema offered by Newell and Simon (1976) and Marr (1982), Sun et al. (2005) proposed that integration across different levels of cognitive model may be necessary for an adequate understanding of cognition and attendant phenomena. Whereas predecessors focused on activity at different levels within individual cognitive agents, Sun et al. (2005) argued that any adequate account of intelligence must consider factors in terms of which intelligence emerges and in terms of which intelligent agents act, set out across four different levels of organization spanning (top to bottom): that accessible to sociological and anthropological inquiry including relationships between individuals and their environments, individual behavior as accessible to psychological inquiry, specialized modules and their assembly into functional whole brains accessible to cognitive science (as traditionally understood), and self-organizing physiochemical systems, i.e., living systems as accessible to fundamental biochemical inquiry (Sun et al. 2005, discussion pp. 619–621).	В исправлении предыдущей схемы, предложенной Newell and Simon (1976) и Marr (1982), Сун и соавт. (2005) предположили, что интеграция на разных уровнях когнитивной модели может быть необходима для адекватного понимания когнитивных и сопутствующих явлений. Принимая во внимание, что предшественники фокусировались на активности на разных уровнях внутри отдельных когнитивных агентов, Сун и соавт. (2005) утверждали, что любой адекватный учет интеллекта должен учитывать факторы, с помощью которых возникает интеллект, и с точки зрения действия интеллектуальных агентов, изложенных на четырех различных уровнях организации (сверху вниз): доступный для социологических и антропологических исследований. включая отношения между людьми и окружающей их средой, индивидуальное поведение как доступное для психологических исследований, специализированные модули и их объединение в функциональные целые мозги, доступные для когнитивной науки (как традиционно понимается), и самоорганизующиеся физико-химические системы, то есть живые системы, доступные для фундаментальных биохимическое исследование (Сан и соавт. 2005, обсуждение на стр. 619–621).
One reason given for Sun et al. (2005) development of this expanded schema is that, regardless of field, it is impractical for a single science to yield complete information at every level of organization at once. Rather, scientists working at one level of organization routinely rely on those at others to account for phenomena in terms outside of focal areas; and, in their discourse, understanding at one level is refined as it is checked against results from other levels, thereby improving the accuracy and predictive	Одна из причин, приведенных Суном и соавт. (2005) разработка этой расширенной схемы заключается в том, что, независимо от области, для одной науки нецелесообразно одновременно получать полную информацию на каждом уровне организации. Скорее, ученые, работающие на одном уровне организации, обычно полагаются на тех, кто работает на других, для объяснения явлений вне основных областей; и в их дискурсе понимание на одном уровне уточняется, поскольку оно

power of all. Sun et al. (2005) contend that we should expect the same dynamics to play out in the cognitive sciences, with models ultimately enriched to represent dynamics at increasingly higher and lower levels of organization in increasingly realistic terms. To increase the psychological realism of social simulations for example, Sun (2012) explicitly calls for the grounding of the social in the cognitive sciences. Most recently, Sun (2018b) calls for the “blending” (p. 245) of cognitive with social models as well, in order that their “integration” results in “tools for more precisely understanding policy implications at both individual and social levels” (p. 240) at the same time.	сверяется с результатами других уровней, тем самым улучшая точность и предсказательную силу всех. Сун и соавт. (2005) утверждают, что мы должны ожидать, что та же динамика будет развиваться в когнитивных науках, в то время как модели, в конечном счете, будут обогащены, чтобы представлять динамику на все более высоких и более низких уровнях организации в более реалистичных терминах. Например, чтобы повысить психологический реализм социальных симуляций, Сун (2012) явно призывает основать социальное в когнитивных науках. Совсем недавно Сун (2018b) призывает к «слиянию» (стр. 245) когнитивных и социальных моделей, чтобы их «интеграция» привела к «инструментам для более точного понимания последствий для политики как на индивидуальном, так и на социальном уровнях» (стр. 240) одновременно.
Specifically on the issue of the purposeful development of integrative models for the information of public policy, Sun (2018b) argues that computational models may benefit policy makers who instead of “relying on speculations” need a more “reliable means for understanding” policy implications (p. 240). Psychologically realistic models of social systems “may be used to predict human performance in organizational settings and to prescribe optimal or nearoptimal cognitive abilities for individuals for specific tasks and organizational structures” (p. 243). Their development “for improved policy making” may take advantage of the “prior validation” of established work from which they are assembled, as demonstrated successes of component models “may be leveraged in validating the overall simulation results” (p. 244). Importantly, this transfer of validity should also extend to policy decisions made on the basis of given results, with the first step being the development of simulations tailored from established and ongoing research to this end.	В частности, по вопросу целенаправленной разработки интегративных моделей для информирования государственной политики, Сун (2018b) утверждает, что вычислительные модели могут принести пользу лицам, определяющим политику, которым вместо «опоры на спекуляции» нужны более «надежные средства для понимания» последствий для политики (стр. 240). Психологически реалистичные модели социальных систем «могут использоваться для прогнозирования человеческой деятельности в организационных условиях и для определения оптимальных или почти оптимальных когнитивных способностей людей для конкретных задач и организационных структур» (стр. 243). Их разработка “в целях совершенствования процесса разработки политики "может опираться на “предварительное подтверждение" установленных результатов работы, из которых они собираются, поскольку продемонстрированные успехи компонентных моделей “могут быть использованы для проверки общих результатов моделирования" (стр. 244). Важно отметить, что эта передача подтверждения должна также распространяться на политические решения, принимаемые на основе заданных результатов, причем первым шагом является разработка симуляций, адаптированных от устоявшихся и текущих исследований с этой целью.
In short, the selective integration of computational models may afford a privileged window on policy implications, and leveraging established programs do so in a reliable and timely manner. The practical issue for the cognitive social scientist becomes, then, understanding how programs at	Короче говоря, избирательная интеграция вычислительных моделей может предоставить привилегированное представление о последствиях для политики, а использование существующих программ делает это надежным и своевременным способом. Таким образом, практическим

different levels of inquiry might fit together and in what terms their integration may take place so as to best inform this process. This is the subject of the next section.	вопросом для когнитивного социолога становится понимание того, как программы на разных уровнях исследования могут сочетаться друг с другом и в каких условиях может происходить их интеграция, чтобы наилучшим образом проинформировать этот процесс. Это тема следующего раздела.
3Possible terms of integration	3 Возможные условия интеграции
To weigh options, we need to account for both human as well as ecological costs due to a given policy or change and compare how they differ over time. To "blend" cognitive with social models to provide a means for this comparison, so that we may more precisely understand the implications of certain practices at individual, social and environmental levels all at once, terms of integration must be identified that facilitate their account in common.	Чтобы взвесить варианты, нам необходимо учесть как человеческие, так и экологические издержки, обусловленные конкретной политикой или изменениями, и сравнить их с течением времени. Чтобы «смешать» когнитивные и социальные модели, для того чтобы обеспечить средства для этого сравнения, чтобы мы могли более точно понять последствия определенных практик на индивидуальном, социальном и экологическом уровнях одновременно, должны быть определены условия интеграции, которые облегчают их учет в общий.
One possibility exists in Karl Friston et al.'s recent work in Markov blankets (cf. Friston 2013; Cockshott and Renaud 2016; Ramstead et al. 2018) providing a framework for the translation between levels of description of human cognition in terms of Friston's "free energy principle" (FEP) (cf. Friston 2010, 2012). What is especially promising about this approach is that it accounts for cognition in terms of and as constrained by natural energetics beginning with Friston's FEP, which for our purposes provides a conceptual bridge between changes in human costs due to policies supporting given institutional arrangements over relatively short time spans (typically intended to reduce costs to some human beings at the expense of their environments, including very often other human beings) and demands on supporting natural environments due those same policies and institutional arrangements over longer time spans.	Одна из возможностей существует в недавней работе Карла Фристона и др. В «Марковских ограждениях» (см. Friston 2013; Cockshott и Renaud 2016; Ramstead et al. 2018), обеспечивающей основу для перевода между уровнями описания человеческого познания в терминах Фристона. «Принцип свободной энергии» (FEP) (ср. Friston 2010, 2012). Что особенно многообещающе в этом подходе, так это то, что он учитывает познание с точки зрения естественной энергетики и ограничивается ею, начиная с ПСЭ Фристона, что для наших целей обеспечивает концептуальный мост между изменениями в человеческих затратах благодаря политике, поддерживающей данные институциональные механизмы в течение относительно короткого времени. промежутки времени (как правило, предназначенные для снижения затрат для некоторых людей за счет их среды обитания, в том числе очень часто других людей) и потребности в поддержке естественной среды из-за той же политики и институциональных механизмов в течение более продолжительных периодов времени.
Friston's free energy principle (FEP) formalizes cognition in terms of the maximization of expected utility, reward or value, through the minimization of prediction error, surprise, or cost, by way of "active inference" which involves maximizing evidence for internal models of the world as informed through ongoing sensory input. Generally, Friston et al. understand that neural structures within an organism work to minimize differences between anticipated ends and perceived results, with future intentions to act modified	Принцип свободной энергии Фристона (ПСЭ) формализует познание с точки зрения максимизации ожидаемой полезности, вознаграждения или ценности посредством минимизации ошибки предсказания, неожиданности или стоимости посредством «активного вывода», который включает в себя максимальное доказательство наличия внутренних моделей мир, как сообщается через постоянный сенсорный вклад. Как правило, Фристон и соавт. понимают, что нейронные

accordingly, and with the aim of this process being to secure the organism's present and future integrity against disintegrative change. The FEP expresses the key relationship between the cognitive agent's perceived and anticipated possible ends in terms of uncertainty, with the agent essentially motivated to avoid surprise. When anticipated ends match perceived results, the surprise is zero. When they do not, surprise demands attention and resources are expended. This is important given a scarcity of resources, and motivates agent psychology generally. Complicated internal models, political philosophies and economic systems are all expressions of cognitive systems operating according to this organizational principle, developing according to the implicit aim of minimizing uncertainty through the proactive organization of self, other and environment at the expense of energies collected and distributed through increasingly complex social arrangements.	структуры в организме работают для минимизации различий между ожидаемыми целями и предполагаемыми результатами, с будущими намерениями действовать соответствующим образом, и с целью этого процесса состоит в том, чтобы обезопасить нынешнюю и будущую целостность организма от дезинтегративных изменений. ПСЭ выражает ключевые отношения между предполагаемыми и ожидаемыми возможными целями когнитивного агента с точки зрения неопределенности, причем агент по существу мотивирован, чтобы избежать неожиданности. Когда ожидаемый конец соответствует ожидаемому результату, неожиданность равна нулю. Когда это не так, неожиданность требует внимания, и ресурсы расходуются. Это важно, учитывая нехватку ресурсов, и мотивирует агентскую психологию в целом. Сложные внутренние модели, политическая философия и экономические системы - все это выражения когнитивных систем, работающих в соответствии с этим организационным принципом, которые развиваются в соответствии с неявной целью минимизации неопределенности посредством упреждающей организации себя, других и окружающей среды за счет энергий, собранных и распределенных посредством все более сложных социальных мероприятий.
Due to the explanatory scope of this program, it presents itself as providing possible terms for the integration of cognitive models required should policy informing simulations of the sort proposed by White (2016) and by Sun (2018b) be realized. Noteworthy in the present context is that Friston's FEP has already been employed in inquiries into cognition at different levels of organization.	В связи с пояснительной областью применения этой программы она представляет собой возможные условия для интеграции когнитивных моделей, требуемых в случае реализации имитирующих политику моделей, предложенных Уайтом (2016) и Сунь (2018b). В нынешнем контексте следует отметить, что ПСЭ Фристона уже использовалась в исследованиях познания на разных уровнях организации.
•Ramstead et al. (2018) use the FEP to characterize cognition in terms of a general theory of dynamic systems consistent with evolutionary systems theory (EST). On this account, cognition is an aspect of living systems which maintain themselves in a limited number of stable states far from thermodynamic equilibrium with their environments by organizing themselves and their environments in such a way as to minimize disorder and surprise at the failure to deliver anticipated results, in effect securing preferred modes of "coupling" with their environments. "Consistent with EST, this propensity to minimize sur-prise is the result of natural selection …self-organizing systems that are able to avoid entropic, internal phasetransitions have been selected over those that could not" (Ramstead et al. 2018, p. 3).	• Рамстед и соавт. (2018) используют ПСЭ для характеристики познания в терминах общей теории динамических систем, соответствующей теории эволюционных систем (ТЭС). В связи с этим познание является одним из аспектов живых систем, которые поддерживают себя в ограниченном числе стабильных состояний, далеких от термодинамического равновесия со своей средой, организуя себя и свою среду таким образом, чтобы свести к минимуму беспорядок и удивление по поводу неспособности достичь ожидаемых результатов, фактически обеспечивая предпочтительные способы "сцепления" со своей средой. «В соответствии с ТЭС, эта склонность к минимизации неожиданности является результатом естественного отбора… самоорганизующиеся системы, которые способны избегать энтропийных, внутренних

	фазовых переходов, были выбраны из тех, которые не могли» (Ramstead et al. 2018, p. 3).
•At the level of social organization in the context of economics, the free energy principle has proven more successful in understanding choice behavior as the minimization of surprise coupled with utility maximization than have other approaches which try to model the same phenomena in terms of utility maximization alone (Schwartenbeck et al. 2015).	• На уровне социальной организации в контексте экономики принцип свободной энергии оказался более успешным для понимания поведения выбора как минимизации неожиданности в сочетании с максимизацией полезности, чем другие подходы, которые пытаются моделировать те же явления с точки зрения максимизации одной полезности (Schwartenbeck et al. 2015).
•At the level of situated cognition, the FEP has recently been deployed in understanding how stress contributes to disease in organisms. Peters et al. (2017) interpret stress according to the free energy principle as “uncertainty” or “surprise” that frustrates the fundamental motivation to minimize entropy. In response, the brain taxes the body system by demanding more energy to rectify the condition and thereby diverting attention away from immediate tasks, with this “allostatic load” resulting in impaired memory, increased markers for cardiovascular disease, and diabetes.	• На уровне когнитивного познания ПСЭ недавно был применен для понимания того, как стресс способствует болезням в организмах. Петерс и соавт. (2017) интерпретируют стресс в соответствии с принципом свободной энергии как «неопределенность» или «неожиданность», которые расстраивают фундаментальную мотивацию минимизации энтропии. В ответ мозг нагружает систему организма, требуя больше энергии для исправления состояния и, таким образом, отвлекая внимание от непосредственных задач, с этой «аллостатической нагрузкой», приводящей к ухудшению памяти, увеличению маркеров сердечно-сосудистых заболеваний и диабета.
•And in neuropsychology for example, Friston’s FEP has been employed in the study of mirror neuron activity in the motor cortex in order to understand how brains switch between perception and action (cf. Shipp et al. 2013), as well as in accounting for reward learning as driven by dopamine excitation or depletion (Fitzgerald et al. 2015; see also Friston et al. 2016).	• И в нейропсихологии, например, ПСЭ Фристона использовался для изучения активности зеркальных нейронов в моторной коре, чтобы понять, как мозг переключается между восприятием и действием (см. Shipp et al. 2013), а также для учета поощрения обучения, обусловленное возбуждением или истощением дофамина (Fitzgerald et al. 2015; см. также Фристон и соавт. 2016).
The next section introduces three different research programs that fall roughly into the evolutionary, social and neurodynamic levels of organization, before briefly sketching how they may all work together to advise public policy in Sect. 5.	В следующем разделе представлены три различные исследовательские программы, которые примерно соответствуют эволюционному, социальному и нейродинамическому уровням организации, а затем кратко рассмотрим, как все они могут работать вместе для выработки рекомендаций по государственной политике в разделе 5.
4 Levels of model	4 Уровни модели
This section reviews three active research programs in cognitive modeling, each at a different level of organization. The first is Peirera et al.’s evolutionary psychological approach employing logic programming to investigate the effects of different expressions of moral agency (apology, forgiveness, preconditions to cooperation, and so on) on group performance over evolutionary time. The second is Sun et al.’s cognitive social science approach focusing on psychologically realistic computational models of	В этом разделе рассматриваются три активные исследовательские программы в области когнитивного моделирования, каждая на своем уровне организации. Первым является эволюционный психологический подход Пейрера и др., Использующий логическое программирование для исследования влияния различных проявлений морального духа (извинения, прощения, предварительных условий для сотрудничества и т. д.) На групповые результаты в течение эволюционного времени.

social intelligence. The third is Tani et al.'s neurodynamic approach grounded on predictive coding and directly demonstrative of Friston's FEP in learning neurorobots. The fifth section then sketches in general terms how these three may be integrated to assess human and natural resource costs of competing policy proposals.	Второй - подход когнитивной социальной науки Сун и соавт., фокусирующийся на психологически реалистичных вычислительных моделях социального интеллекта. Третий - нейродинамический подход Тани и соавт., Основанный на прогнозирующем кодировании и прямо демонстрирующий ПСЭ Фристона в обучении нейроборотов. Затем в пятом разделе в общих чертах описывается, как эти три программы могут быть объединены для оценки затрат на людские и природные ресурсы конкурирующих политических предложений.
4.1Pereira's evolutionary game theory	4.1 Эволюционная теория игр Перейры
L. M. Pereira et al. use logic programming and evolution ary game theory to model the capacity for individual agents to make moral decisions through abduction, either in reaction to contextual cues or through purposeful deliberation over points of interest, and from this basis have worked on understanding the roles of intention recognition, commitment, apology, forgiveness, revenge, ostracism, and guilt in cooperative collectives of similarly endowed individuals. The aim of this research is to better understand the emergence of cooperation as supported by cognitive mechanisms which thereby stabilize social orders over evolutionary time scales, so that this understanding may both inform human practice today, and so that such capacities may be interred in future robotic agents free to act within future human communities tomorrow. Thus, one strong theme running through Pereira et al.'s work has been the need to bridge individual with collective "realms" toward the goal of understanding just how individual agents act from the basis of one in the furtherance of the other (cf. Pereira and Saptawijaya 2015, 2016; Saptawijaya and Pereira 2018; also Han and Pereira 2018).	Л. М. Перейра и соавт. используют логическое программирование и эволюционную теорию игр, чтобы смоделировать способность отдельных агентов принимать моральные решения посредством абдукции, либо в ответ на контекстные сигналы, либо посредством целенаправленного обсуждения вопросов, представляющих интерес, и на этой основе работали над пониманием роли распознавания намерений приверженность, извинения, прощение, месть, остракизм и вина в кооперативных коллективах одаренных людей. Цель этого исследования - лучше понять возникновение сотрудничества, поддерживаемое когнитивными механизмами, которые тем самым стабилизируют социальные порядки в эволюционных временных масштабах, так что это понимание может как информатировать человеческую практику сегодня, так и такие возможности могут быть использованы в будущем роботизированными агентами, которые могут свободно действовать в будущих человеческих сообществах завтра. Таким образом, одной сильной темой, пронизывающей работу Перейры и соавт., была необходимость связать индивида с коллективными «сферами» для достижения цели понимания того, как отдельные агенты действуют от основания одного в поддержке другого (ср. Pereira and Saptawijaya 2015, 2016; Saptawijaya и Pereira 2018; также Han и Pereira 2018).
Pereira et al. account for native agent motivation to best available ends in terms of abduction. Historically, abduction has been variably given, depending on researcher and context. For Peirce—the inventor of the concept—abduction was variably characterized as well, depending on at which stage of his life one were to have asked him about it. According to a mature view, abduction is a natural tendency to discovery of truth, a "guessing instinct" through which (typically more successful than not)	Перейра и соавт. учитывают мотивацию родного агента к наилучшим доступным целям с точки зрения абдукции. Исторически абдукция была разной, в зависимости от исследователя и контекста. Для Пирса - изобретателя концепции - абдукция также характеризовалось по-разному, в зависимости от того, на каком этапе его жизни нужно было спросить его об этом. Согласно зрелой точке зрения, абдукция является естественной тенденцией к раскрытию истины, «инстинктом

hypotheses are created and provisionally adopted as they are then tested through induction and clarified through deduction before contributing to further creative abduction (cf. Paavola 2006, discussion Chap. 4; also Aliseda 2006; Gabbay and Woods 2005; and Magnani 2017; see also Simon 1977, for an early view on abduction encoded as a computer program). In Pereira et al.'s models, abduction is a matter of determining a set of actions that satisfy goal conditions while maintaining personal integrity. These are iterated as "abducibles" and are evaluated by agents using doctrines of double and triple effect, utility functions, and counterfactuals for example (cf. Han and Pereira 2013; Han et al. 2015; Pereira and Saptawijaya 2017). The ethical norms of a group evolve as agents organize around increasingly ideal solutions afforded by increased agent-level capacities to cooperate, pursuing strategies for higher group-level payoffs of which agents share.	угадывания», посредством которого (как правило, более успешные, чем нет) гипотезы создаются и временно принимаются, поскольку они затем проверяются путем индукции и разъясняются путем дедукции, прежде чем вносить вклад в дальнейшую творческую абдукцию (см. Paavola 2006, обсуждение главы 4; также Aliseda 2006; Gabbay and Woods 2005; и Magnani 2017; см. также Simon 1977, для раннего взгляда на абдукцию, закодированная как компьютерная программа). В моделях Перейры и др. абдукция является вопросом определения набора действий, которые удовлетворяют целевым условиям при сохранении личной неприкосновенности. Они повторяются как «выводимые» и оцениваются агентами, используя, например, доктрины двойного и тройного эффекта, функции полезности и контрфактуальность (см. Han и Pereira 2013; Han et al. 2015; Pereira and Saptawijaya 2017). Этические нормы группы развиваются по мере того, как агенты организуются вокруг все более и более идеальных решений, которые обеспечиваются за счет повышения способности сотрудничать на уровне агентов, следуя стратегиям для более высоких выплат на уровне групп, которые разделяют агенты.

On Pereira et al.'s model, an agent's prior deliberate decisions to act in given situations are retained, and these can be employed reactively in similar future situations without the need to compute them again (cf. Saptawijaya and Pereira 2013; also Saptawijaya and Pereira 2018). Moreover, this "tabling" technique opens prior decisions to comparison between agents and allows for different agents to inform each other about differently determined optimal actions in different ways (cf. Pereira et al. 2013). Agents are also able to recognize intentions, to assess relative commitments to goals, to cooperate with each other where projected payoffs are better, and learn to coordinate actions only with those also prone to cooperation. With other prosocial capacities, such as the abilities to issue and to accept apologies, along with capacities to adjust internal commitments to future cooperation, agents otherwise marginalized by past mistakes or bad information are able to again contribute to cooperative endeavors. As a result, these agents learn to share in the mutual benefits of goals unassailable to the isolated individual and realizable only through more complex social interaction, confirming the precedence of prosocial capacities in the evolution of ethics (cf. Han et al. 2011, 2012, 2013; Han 2013).	В модели Перейры и соавторов прежние преднамеренные решения агента действовать в данных ситуациях сохраняются, и они могут использоваться реактивно в аналогичных будущих ситуациях без необходимости их повторного вычисления (см. Saptawijaya и Pereira 2013; также Saptawijaya и Перейра 2018). Более того, этот метод «табулирования» открывает предварительные решения для сравнения между агентами и позволяет разным агентам по-разному информировать друг друга о по-разному определенных оптимальных действиях (см. Pereira et al. 2013). Агенты также могут распознавать намерения, оценивать относительные обязательства по отношению к целям, сотрудничать друг с другом в тех случаях, когда прогнозируемые выплаты лучше, и учиться координировать действия только с теми, кто склонен к сотрудничеству. Благодаря другим способностям просоциальности, таким как способность выдавать и принимать извинения, а также способность корректировать внутренние обязательства по будущему сотрудничеству, агенты, в противном случае оказавшиеся в стороне от прошлых ошибок или недостоверной информации, могут снова внести свой вклад в совместные усилия. В результате эти агенты учатся разделять взаимные выгоды от целей, недоступных для изолированного индивида и достижимых только через более сложное социальное взаимодействие, подтверждая приоритет просоциальных способностей в эволюции этики (см. Han et al. 2011, 2012 2013; Han 2013).
Pereira et al.'s research pursues a bottom-up explanation for the emergence of morality over evolutionary time. Their work confirms that agents with native capacities to better cooperate outperform those without, both in pairwise situations and in common good settings (Han et al. 2017; Martinez-Vaquero et al. 2015, 2017). It leads Pereira et al. to conclude that evolved cognitive capacities facilitating cooperation induce the emergence of what we recognize as morality in human populations (as opposed to stable cultural practices as systems of ethics understood as rules and institutions coming first, inducing cooperation instead, cf. Pereira and Saptawijaya 2015; Han and Pereira 2018). This result has important implications for moral education in youth for example, illustrating at once how research at this high level of organization can inform research at lower levels of organization in constructive ways. Moreover, it confirms that lower-level dynamics are critical to understanding social norms and commensurate public policies.	Исследование Перейры и соавт. преследует восходящее объяснение появления морали с течением времени. Их работа подтверждает, что агенты с собственными способностями к лучшему сотрудничеству превосходят тех, у кого нет, как в парных ситуациях, так и в общих хороших условиях (Han et al. 2017; Martinez-Vaquero et al. 2015, 2017). Это приводит Перейра и соавт. сделать вывод о том, что развитые познавательные способности, способствующие сотрудничеству, вызывают появление того, что мы признаем нравственностью в человеческом населении (в отличие от устойчивых культурных практик, поскольку системы этики, понимаемые как правила и институты, стоят на первом месте, стимулируя сотрудничество, см. Перейра и Саптавиджая 2015; Хан и Перейра 2018). Этот результат имеет важные последствия для нравственного воспитания молодежи, например, сразу демонстрируя, как исследования на этом высоком уровне организации

	могут конструктивно влиять на исследования на более низких уровнях организации. Более того, это подтверждает, что динамика более низкого уровня имеет решающее значение для понимания социальных норм и соразмерной государственной политики.
4.2Sun's cognitive social sciences	4.2. Когнитивные социальные науки Суна
The upshot of cognitive modeling or computational models "in the broad sense of the term" according to Sun et al. (2005) is that these models serve as operational frameworks—"broad, generic theories of cognition" (Sun et al. 2005, p. 616)—for the structured interpretation of "a vast amount of data" generated by the cognitive and social sciences (Sun et al. 2005, p. 615). Where Pereira et al. focus on the consequences of routine action over generations of psychologically simple moral agents, Sun et al. focus on an equally vast area, the psychological processes that render such actions, rules expressing them and further their revision through the interplay of different specialized modules of information processing constitutive of more psychologically realistic learning agents. Bridging the second and third of four levels of cognitive model as iterated at the beginning of the second section of this paper, much of Sun's research focuses on multi-agent systems, social interaction, and prosocial motivation including aspects overlapping Pereira et al.'s work at higher levels of organization such as social stabilizing capacities involving intention recognition. With a resolution on cognitive mechanisms beneath the lower bounds of their evolutionary framework, however, Sun's trademark Clarion computational architecture also resolves cognition at levels of organization overlapping in part with Tani et al.'s focus on fundamental neurodynamics, for example attending to social autonomy and emotion among other psychological constructs (cf. Sun 2002, 2009, 2013, 2016, 2017, 2018a, b; Sun and Naveh 2004; Sun et al. 2016).	Результатом когнитивного моделирования или вычислительных моделей «в широком смысле этого слова» в соответствии с Сун и соавт. (2005) заключается в том, что эти модели служат в качестве операционных рамок - «широких, общих теорий познания» (Sun et al. 2005, p. 616) - для структурированной интерпретации «огромного количества данных», генерируемых когнитивными и социальными науками (Сан и соавт. 2005, с. 615). Где Перейра и соавт. сосредоточиваются на последствиях рутинных действий над поколениями психологически простых моральных агентов, Сун и соавт. сосредоточили внимание на такой же обширной области, психологических процессах, которые делают такие действия, правилах, выражающих их, и дальнейшем их пересмотре посредством взаимодействия различных специализированных модулей обработки информации, составляющих более психологически реалистичных обучающих агентов. Соединяя второй и третий из четырех уровней когнитивной модели, как повторено в начале второго раздела этой статьи, большая часть исследований Суна фокусируется на мультиагентных системах, социальном взаимодействии и просоциальной мотивации, включая аспекты, перекрывающие работу Перейру и соавт. на более высоких уровнях организации, таких как социальные стабилизирующие способности, предполагающие признание намерений. Однако благодаря разрешению когнитивных механизмов ниже нижних границ их эволюционной структуры вычислительная архитектура Clarion торговой марки Sun также разрешает познание на уровнях организации, частично совпадающих с акцентом Тани и соавт. На фундаментальной нейродинамике, например на социальной автономии. и эмоции среди других психологических конструктов (ср. Сун 2002, 2009, 2013, 2016, 2017, 2018a, b; Sun and Naveh 2004; Sun et al. 2016).
Consider, for example, the relationship between Sun's research program and Pereira's in greater detail. Clarion is motivated by 11 primary drives, of which many correspond to native capacities to cooperate on Pereria et al.'s	Рассмотрим, например, связь между исследовательской программой Суна и Перейрой более подробно. Clarion мотивируется 11 первичными стимулами, многие из которых соответствуют собственным

model. For example, “similance”—the drive for one to identify with, and to emulate others—along with other primary drives including those to affiliate with others and to belong to groups, to avoid harmful situations, to resist control and to ensure that one’s self and others are treated fairly, all work together to demonstrate a more detailed model of moral agent psychology unnecessary at Pereira et al.’s scale of evolutionary game theory (see Sun 2017, Table 1, p. 6, for a most recent summary of drives in Clarion). Clarion’s psychological realism sets it apart from Pereira et al.’s model agents in other ways as well. For example, Clarion has been assessed for consciousness alongside competing architectures, and found to represent aspects of consciousness including qualia (Gok and Sayan 2012). However, like Pereira et al.’s model agents, Clarion does not aim to replicate human biological cognition or to capture the principles of its self-organization, and rather works at the level above physiological processes on Sun et al.’s four-tier scheme, at the level of psychological processes instead.	способностям сотрудничать по модели Перерии и др. Например, «подобие» - стремление человека идентифицировать себя и подражать другим - наряду с другими основными побуждениями, в том числе связанными с другими и принадлежать к группам, чтобы избежать вредных ситуаций, противостоять контролю и обеспечить к себе и другим относятся справедливо, все работают вместе, чтобы продемонстрировать более детальную модель психологии морального агента, ненужную в шкале эволюционной теории игр Перейры и др. (см. Сун 2017, Table 1, p. 6, для последнего резюме дисков в Clarion). Психологический реализм Clarion отличает его от модельных агентов Перейры и других также и в других отношениях. Например, Clarion был оценен на предмет сознания вместе с конкурирующими архитектурами, и было обнаружено, что он представляет аспекты сознания, включая qualia (Gok and Sayan 2012). Однако, подобно модельным агентам Перейры и соавторов, Кларион не ставит своей целью копирование биологического познания человека или улавливание принципов его самоорганизации, а скорее работает на уровне выше физиологических процессов на четырехстороннем опыте Сун с соавторами. Ярусная схема, на уровне психологических процессов.
At the center of Sun’s study of psychological processes is a “causal nexus” of activity between the implicit and explicit modes of information processing characteristic of hybrid systems of which Clarion is an example. Hybrid systems represent higher and lower cognitive functions in different ways, and these can interact with each other in upand downstream processes. Clarion consists of a number of hybrid subsystems (cf. Sun 2002) whose bottom levels mediate routine action and encode regularities from which top-level rules are extracted and which are then applied topdown in the direction of future action (Sun et al. 2001; Sun 2016). As the model agent learns (bottom-up) to autonomously specify and modulate goals (which can also be learned top-down), relatively stable constructs within the cognitive architecture amount to what we recognize as “personality” in human beings (cf. Sun and Wilson 2014). Stable personalities make for predictable intentions, which Sun and Pereira capture in ways that complement one another. Sun et al.’s view complements that established by Pereira et al. by extending insight into those cognitive capacities which contribute to conventions and institutions grounding lasting social–political systems (and so mechanisms of artificial selection influencing human evolution at the same	В центре изучения психологических процессов Суна находится «причинно-следственная связь» активности между неявным и явным способами обработки информации, характерными для гибридных систем, примером которых является Clarion. Гибридные системы по-разному представляют высшие и низшие когнитивные функции, и они могут взаимодействовать друг с другом в процессах восходящего и нисходящего потоков. Clarion состоит из ряда гибридных подсистем (ср. Сун 2002), нижние уровни которых опосредуют рутинные действия и кодируют закономерности, из которых извлекаются правила верхнего уровня и которые затем применяются сверху вниз в направлении будущих действий (Сан и соавт. 2001; Вс 2016). Когда модельный агент учится (снизу вверх) автономно определять и модулировать цели (которые также можно изучать сверху вниз), относительно стабильные конструкции в рамках когнитивной архитектуры составляют то, что мы распознаем как «личность» в людях (ср. Сун и Уилсон 2014). Стабильные личности создают предсказуемые намерения, которые Сун и Перейра улавливают так, что дополняют друг друга. Взгляд Суна и соавт. дополняет точку зрения Перейра и соавт. расширяя понимание

time) as these, at a finer grain of analysis, are strengthened into agent-specific personalities in a context-dependent manner. Moreover, occupying this middle space between biological constitution and cultural realization, Sun's program affords an inroad into larger systems for the influence of cognitive processes resolved by models at lower levels of organization designed to capture the dynamics of such personality formation internal to the individual agent, itself.	тех познавательных способностей, которые способствуют конвенциям и институтам, создающим основу для устойчивых социально-политических систем (и, таким образом, механизмов искусственного отбора, влияющих одновременно на эволюцию человека), так как они, при более глубоком анализе, превращаются в специфические для агентов личности в зависимости от контекста. Более того, занимая это промежуточное пространство между биологической конституцией и культурной реализацией, программа Sun позволяет проникнуть в более крупные системы влияния когнитивных процессов, разрешаемых моделями на более низких уровнях организации, предназначенных для отражения динамики такого формирования личности, внутренней по отношению к индивидуальному агенту, сам.
4.3Tani's fundamental neurodynamics	4.3 Фундаментальная нейродинамика Тани
Predictive coding along with active inference and the FEP constitute an important approach to understanding how cognition works at different levels of organization, serving as a broad framework according to which vast amounts of data from cognitive and social sciences can be interpreted. At the level of neurodynamics, its efficacy is confirmed by how well Tani et al.'s computational models demonstrating these principles articulate biological cognition in neurorobots. This section reviews Tani et al.'s program, and the next section sketches how the principles underwriting this research may be extended to cognitive models at higher levels of organization in their integration toward platforms designed to inform public policy.	Прогнозирующее кодирование наряду с активным умозаключением и СЭФ представляют собой важный подход к пониманию того, как познание работает на различных уровнях организации, и служит широкой основой, в соответствии с которой могут интерпретироваться огромные объемы данных из когнитивных и социальных наук. На уровне нейродинамики ее эффективность подтверждается тем, насколько хорошо вычислительные модели Тани и соавт., демонстрирующие эти принципы, определяют биологическое познание у нейроборотов. В этом разделе рассматривается программа Тани и соавт., а в следующем разделе показано, как принципы, лежащие в основе этого исследования, могут быть распространены на когнитивные модели на более высоких уровнях организации при их интеграции в платформы, предназначенные для информирования государственной политики.
Recalling the brush with dynamic systems theory and predictive coding in Sect. 3, on Tani et al.'s program a learning agent develops an "internal model" of the world as a set of self-organized dynamic attractors (Tani 1996; Tani and Nolfi 1999) toward which future actions aim. These aims are then challenged in the conflictive interaction between topdown and bottom-up processing streams (note the parallel with Sun's "causal nexus") as the perceived world deviates from model projections, resulting in an unstable "critical" state followed by the effortful return to stable coherency with the perceived reality as this internal model is recomposed (cf. Tani 2007). This process is repeated over time in various environments as the agent learns to	Вспоминая куст теории динамических систем и прогнозирующего кодирования в разд. 3, в программе Тани и др. Учебный агент разрабатывает «внутреннюю модель» мира как набор самоорганизующихся динамических аттракторов (Тани 1996; Тани и Нолфи 1999), на которые нацелены будущие действия. Затем эти цели ставятся под сомнение в конфликтном взаимодействии между потоками обработки сверху вниз и снизу вверх (обратите внимание на параллель с «причинной связью» Суна), поскольку воспринимаемый мир отклоняется от проекций модели, что приводит к нестабильному «критическому» состоянию, за которым следует возврат усилий.

achieve different goals, and the result is an artificial embodiment of the principles that account for similar processes in biological cognition (White and Tani 2016, 2017; Tani and White 2017).	устойчивой согласованности с воспринимаемой реальностью, поскольку эта внутренняя модель перекомпонована (ср. Тани 2007). Этот процесс повторяется с течением времени в различных средах, когда агент учится достигать разных целей, и в результате получается искусственное воплощение принципов, которые учитывают сходные процессы в биологическом познании (Уайт и Тани 2016, 2017; Тани и Уайт 2017).
For example, employing a relatively simple architecture using RNNs tuned to different timescales, with the lower level at a shorter timescale sensitive to rapid changes in the environment, and the higher level at a slower timescale able to extract longer-standing patterns from the same input (reflecting its predictive coding framework), Nishimoto and Tani (2009) demonstrated the development of a stable functional hierarchy whereby primitive behaviors that develop early on in the lower levels are composed into more complicated action routines in the higher level as the agent learns to achieve increasingly challenging goals during later stages, corresponding to Piaget's constructivist developmental psychology (cf. Piaget 1954). Namikawa et al. (2011) further relate these results to the developmental process of the dynamical hierarchy involving the prefrontal cortex, supplementary motor area and primary motor cortex in human beings during spontaneous composition of complex actions from primitives, as in both computational and biological systems the prefrontal areas develop similarly and—depending on the conditions of this development—deliver similar patterns of behavior.	Например, использование относительно простой архитектуры с использованием RNN, настроенных на разные временные масштабы, с более низким уровнем в более коротком временном масштабе, чувствительным к быстрым изменениям в окружающей среде, и более высоким уровнем в более медленном временном масштабе, способным извлекать более длительные шаблоны из одного и того же входа. (отражая структуру прогнозирующего кодирования), Нишимото и Тани (2009) продемонстрировали развитие стабильной функциональной иерархии, в которой примитивное поведение, которое развивается на ранних этапах на более низких уровнях, состоит из более сложных процедур действий на более высоком уровне, когда агент учится достигать все более сложные задачи на более поздних этапах, соответствующие конструктивистской психологии развития Пиаже (ср. Пиаже, 1954). Намикава и соавт. (2011) далее связывают эти результаты с процессом развития динамической иерархии с участием префронтальной коры, дополнительной моторной области и первичной моторной коры у людей во время спонтанной композиции сложных действий от примитивов, как в вычислительных, так и биологических системах развиваются префронтальные области аналогично и - в зависимости от условий этого развития - обеспечивают сходные модели поведения.
Noteworthy is that Tani's models are not "hybrid" like Sun's, as higher and lower levels share the same metric space, i.e., they do not represent information in different ways, but rather find different patterns in different aspects of ongoing information processing in the same ways. In this way, Tani et al.'s approach to cognitive modeling is able to complement investigations undertaken on Sun et al.'s psychological approach in terms of biologically plausible—rather than psychologically plausible—dynamics. For instance, complementing Sun's (2013) account of creativity due to subsymbolic dynamics in hybrid models, Tani et al. have also investigated how actions are learned and why novel actions are composed. In a social	Следует отметить, что модели Тани не являются «гибридными», как модели Сун, поскольку более высокие и более низкие уровни разделяют одно и то же метрическое пространство, то есть они не представляют информацию по-разному, а скорее находят разные закономерности в разных аспектах текущей обработки информации в одном и том же пути. Таким образом, подход Тани и др. К когнитивному моделированию может дополнить исследования психологического подхода Сун и др. С точки зрения биологически правдоподобной, а не психологически правдоподобной динамики. Например, дополняя учет креативности Сун (2013) благодаря подсимболической динамике в

situation, Ito and Tani (2004) employed a mirror neuron model to investigate how a complex action routine can be encoded as a single “chunk” when agent/environment dynamics are predictable, and how the resulting single seamless operation can then be resegmented into constitutive primitives through backpropagated prediction error when input proves unpredictable, with these pieces then autonomously recomposed into new patterns as the system attempts to restore upand downstream coherency with perceived reality through novel action in response. With a robot motivated by this model to coordinate with a human subject, and with the human simultaneously attempting the same, Tani et al. learned that even small perturbations in the robot’s actions could cause confusion in human subjects while the subjects were becoming accustomed to the robot’s repertoire of learned action sequences (and vice versa). As a result, turn taking (with either robot or human subject leading action sequences) became prevalent during this mutual learning period, a fact that Tani and Ito interpreted as mutually initiated in response to the breakdown of higher-level intentional constructs (or “criticality”) in both human and robot partners.	гибридных моделях, Тани и соавт. также исследовали, как изучаются действия и почему составляются новые действия. В социальной ситуации Ито и Тани (2004) использовали модель зеркальных нейронов, чтобы исследовать, как сложная подпрограмма действия может быть закодирована как единый «кусок», когда динамика агента / среды предсказуема, и как тогда может быть получена единственная бесшовная операция. Сегментируется на конститутивные примитивы с помощью ошибки предсказания с обратным распространением, когда входные данные оказываются непредсказуемыми, а затем эти фрагменты автономно перекомпоновываются в новые шаблоны, когда система пытается восстановить согласованность в восходящем и нисходящем направлениях с воспринимаемой реальностью посредством новых действий в ответ. С роботом, мотивированным этой моделью для координации с человеком, и с человеком, одновременно пытающимся сделать то же самое, Тани и соавт. узнал, что даже небольшие возмущения в действиях робота могут вызвать замешательство у людей, в то время как субъекты привыкли к репертуару роботов с изученными последовательностями действий (и наоборот). В результате, в этот период взаимного обучения стал преобладать ход ходов (с ведущими последовательностями действий либо с роботом, либо с человеком), факт, который Тани и Ито истолковали как взаимно инициированные в ответ на разрушение намеренных конструкций более высокого уровня (или «критичность»). Как у людей, так и у партнеров-роботов.
Murata et al. (2014) further investigated the proactive coordination of one’s own actions with another’s predictable action sequences as opposed to the reactive dynamics which come into play as a predicted action sequence is in error, with interacting agents each aiming to restore coherency between upand downstream processes, echoing again Sun’s focus on this nexus of activity. Murata et al. (2017) extends these results by developing a robot that integrates external with internal sources of information in the continuous performance of sensory-dependent and sensory-independent tasks, thereby operating both online (open to influence from the outside) and off-line (operating according to internal determinations) as task processing demands. Together, we can begin to see how cooperative interaction is motivated by shared neurodynamics, on this portrait as agents variably integrate self and other information online during coordinated interaction, pursuing best prediction from the top-down given bottom-up cues also recalling Pereira et	Мурата и соавт. (2014) дополнительно исследовали проактивную координацию своих собственных действий с предсказуемыми последовательностями действий другого в противоположность реактивной динамике, которая вступает в игру, когда предсказанная последовательность действий ошибочна, причем взаимодействующие агенты стремятся восстановить согласованность между восходящими и нисходящими процессами, отражая снова внимание Суна на этой связке активности. Мурата и соавт. (2017) расширяет эти результаты, разрабатывая робота, который интегрирует внешние с внутренними источниками информации для непрерывного выполнения сенсорно-зависимых и сенсорно-независимых задач, тем самым работая как в режиме онлайн (открытый для влияния извне), так и в автономном режиме (режим работы согласно внутренним определениям) как требования к обработке задачи. Вместе мы можем начать видеть, как

al.’s intention recognition and abducible ends at a much finer grain of analysis.	кооперативное взаимодействие мотивируется общей нейродинамикой, на этом портрете, когда агенты по-разному интегрируют себя и другую информацию онлайн во время скоординированного взаимодействия, следуя наилучшему прогнозированию из нисходящих данных восходящих сигналов, также ссылаясь на Перейру и др. Признание намерений и выводимость заканчиваются гораздо более тонким анализом.
In summary, Tani et al.’s research affords insight into aspects of the human condition that cannot be realistically represented at higher levels of organization, for example into the dynamic origins of agent autonomy. Tani et al.’s model agents learn action primitives during entrainment with their object environments. These learned primitives are then recomposed in response to changing conditions to align information processing streams and maintain a stable internal world model (Tani 2016, see Chap. 8, Sect. 4). The point here is twofold. For one thing, the primitives employed in novel action composition are not freely chosen, but are limited to those already learned. For another, impetus to recompose complex action routines emerges also not as a matter of choice, but rather in response to changes as the project model deviates from perceived reality. On Tani’s account of these dynamics, an agent may feel as if he or she, or it, is radically free to compose novel intentions ex nihilo, as if free to do anything, but this is only due to the lack of access to the processes underlying the composition of actions (see Tani 2016, Chap. 10 for extensive discussion). Given this access, computational models of higher levels of organization may demonstrate how free action may be directed by public policy designed to optimize social conditions for maximal human creativity, cognizant of stressors which, when nearing certain thresholds, may increase rather than decrease adaptability to changing environmental conditions, perhaps by expanding the bounds of social cohesion by encouraging the development of capacities for coordination, for example.	Таким образом, исследования Тани и соавт. дают представление об аспектах состояния человека, которые невозможно реально представить на более высоких уровнях организации, например, о динамическом происхождении агентской автономии. Модельные агенты Тани и соавт. Изучают примитивы действия во время увлечения их объектной средой. Затем эти изученные примитивы перекомпоновываются в ответ на изменяющиеся условия для выравнивания потоков обработки информации и поддержания стабильной модели внутреннего мира (Тани 2016, см. Главу 8, раздел 4). Дело здесь двоякое. С одной стороны, примитивы, используемые в новой композиции действия, не выбираются свободно, а ограничиваются уже изученными. С другой стороны, побуждение к перекомпоновке сложных процедур действий возникает также не по выбору, а скорее в ответ на изменения, когда модель проекта отклоняется от воспринимаемой реальности. С точки зрения Тани об этой динамике, агент может чувствовать, что он или она, или оно, радикально свободны в создании новых намерений ex nihilo, как будто они свободны что-либо делать, но это только из-за отсутствия доступа к процессам. основополагающий состав действий (подробное обсуждение см. в Тани 2016, глава 10). При таком доступе вычислительные модели более высоких уровней организации могут продемонстрировать, как свободная деятельность может быть направлена государственной политикой, направленной на оптимизацию социальных условий для максимального творческого потенциала человека, осознавая факторы стресса, которые при приближении к определенным порогам могут увеличивать, а не снижать приспособляемость к изменениям условия окружающей среды, возможно, путем расширения границ социальной сплоченности, например, путем поощрения развития потенциала для координации.
5 Discussion	5 Обсуждение
Briefly consider how a coordinated development of the three programs	Кратко рассмотрим, как может сыграть скоординированное развитие

reviewed above into a single platform for the information of public policy might play out. Through the extension of fundamental insights from the free energy principle and predictive coding into Sun et al.'s models of social cognition, simulations should prescribe that agents entertain relationships only with those others intent on actions contributing to the minimization of uncertainty. As we extend the results of these models into Pereira et al.'s research, we should find stable arrangements potentiated, arrived at and maintained through the institutions of apology and forgiveness, as well as through promise keeping and transparency of intention. In holding social systems together, we may identify such routine cognitive agency as virtuous, and their contraries vicious. Through simulation of critical periods at this level, systematic reconciliation of the vicious with the virtuous might be recommended, and global agreement protocols permitting a systematic transition toward more stable, more cooperative, less exploitive arrangements may result without the violence that had punctuated historical transitions through similarly critical periods.	трех программ, рассмотренных выше, в единую платформу для информирования о государственной политике. Благодаря расширению фундаментального понимания принципа свободной энергии и прогнозирующего кодирования в моделях социального познания в Сун и соавт. моделирование должно предписывать, что агенты поддерживают отношения только с теми, кто намеревается действовать, способствуя минимизации неопределенности. По мере того, как мы распространяем результаты этих моделей на исследования Перейры и соавт., мы должны находить стабильные договоренности, которые можно укреплять, достигать и поддерживать через институты извинений и прощения, а также благодаря выполнению обещаний и прозрачности намерений. Удерживая социальные системы вместе, мы можем определить такие рутинные когнитивные функции как добродетельные, а их противоречия порочные. Посредством моделирования критических периодов на этом уровне может быть рекомендовано систематическое примирение порочных с добродетельными, и протоколы глобальных соглашений, позволяющие систематический переход к более стабильным, более кооперативным, менее эксплуатирующим соглашениям, могут привести к насилию, которое прерывало исторические переходы через аналогичные критические периоды.
These guidelines may then be passed down through Sun et al.'s level of social agency at the level of rules, for example that agents should act to stabilize expectations within parameters conducive to cooperation and should not engage in deceit, withholding or manipulating information for selfish gain. Simulations at this level should deliver advice on how individuals might best reconcile prior understandings with that understanding necessary to motivate personal changes, facilitating social transitions to institutional arrangements through which goal conditions are realized. Selected cases may then be passed down to the level of internal dynamics as revealed through Tani et al.'s research, and here we may gain some insight into the context-dependent stress that an agent may experience during a given transition along with possible strategies for creatively turning this stress forward into constructive social contributions. With this information, people may be able to take steps to not only minimize the stress of change, but also to maximize the potential to develop healthy responses to it, responses that minimize uncertainty going forward and that at the same time maximize their	Эти руководящие принципы могут затем передаваться через уровень социальной активности Сун и соавт. На уровне правил, например, что агенты должны действовать для стабилизации ожиданий в рамках параметров, способствующих сотрудничеству, и не должны участвовать в обмане, утаивании или манипулировании информацией для корыстной выгоды. Моделирование на этом уровне должно дать совет о том, как люди могут наилучшим образом согласовать предшествующее понимание с этим пониманием, необходимым для мотивации личных изменений, облегчая социальные переходы к институциональным механизмам, посредством которых реализуются целевые условия. Затем отдельные случаи могут быть переданы до уровня внутренней динамики, как показано в исследовании Тани и соавт., и здесь мы можем получить некоторое представление о зависимом от контекста стрессе, который может испытать агент во время данного перехода, наряду с возможными стратегиями для творческого превращения этого стресса в конструктивный социальный вклад.Обладая этой

own free agency, to determine for themselves how they may contribute to pro-social transitions in the interests of justice and the good life in general. These possible solutions may then be passed upward to Sun's and then to Pereira's levels of analysis, therein tested to see how they contribute to the overall stability of the systems that they aim to affect, and then to see if they should inform general principles over the generations that may be necessary to see these affects optimally realized.	информацией, люди могут предпринять шаги, направленные не только на минимизацию стресса от перемен, но и на максимизацию потенциала для выработки здоровых реакций на них, реакций, которые минимизируют неопределенность в будущем и в то же время максимизируют их собственную свободу воли, чтобы определить для себя, как они могут способствовать просоциальным переходам в интересах справедливости и хорошей жизни в целом. Затем эти возможные решения могут быть переданы на уровень анализа Сун, а затем на уровень анализа Перейры, где они проверяются, чтобы увидеть, как они способствуют общей стабильности систем, на которые они стремятся воздействовать, а затем посмотреть, должны ли они информировать общие принципы в течение поколений, которые могут быть необходимы для оптимальной реализации этих воздействий.
From here, simulations may continue, be refined, or restarted using different parameters for evaluation of possible ends given different starting points, for example those representing possible crisis, and a science of suites of simulations such as the one sketched above may establish itself. With AI technology developed in this way, people may eventually be able to choose the world in terms of which they would like to live, and use these simulations to help to plot how to get there, openly, responsibly, at the level of the individual agent in real time and in constant view of the implications of one's actions for broader society, as well as for civilization as a whole.	С этого момента моделирование может продолжаться, уточняться или возобновляться с использованием различных параметров для оценки возможных целей при различных исходных точках, например тех, которые представляют возможный кризис, и может возникнуть наука о множествах имитаций, подобных описанной выше. С помощью технологии искусственного интеллекта, разработанной таким образом, люди могут в конечном итоге выбрать мир, в котором они хотели бы жить, и использовать эти симуляции, чтобы помочь построить план, как добраться туда, открыто, ответственно, на уровне отдельного агента в реальном времени и в постоянном представлении о последствиях своих действий для более широкого общества, а также для цивилизации в целом.
Again, one upshot of this approach is that it takes advantage of increasing momentum in the integration of the cognitive with the social sciences. By beginning with ongoing research programs, the present proposal avoids some costs of development, at the same time leveraging validation of established models to facilitate acceptance of the use of such tools in the information of public policy going forward. Some further upshots of the current approach to integrate across select levels of organization are that computational costs may be lowered relative to more complex simulations intended for other purposes, and that platforms may be more rapidly developed beginning with existing research programs than by developing such predictive simulations ab initio. All together, these advantages go a long way to putting informative simulations within reach, such that, reinforcing Sun (2018b) on this point,	Опять же, одним из результатов этого подхода является то, что он использует преимущества увеличения импульса в интеграции когнитивных с социальными науками. Начинаясь с текущих исследовательских программ, настоящее предложение позволяет избежать некоторых затрат на разработку, в то же время усиливая проверку установленных моделей, чтобы облегчить принятие использования таких инструментов в информации о государственной политике в будущем. Некоторые дальнейшие результаты нынешнего подхода к интеграции на отдельных уровнях организации заключаются в том, что вычислительные затраты могут быть снижены относительно более сложных симуляций, предназначенных для других целей, и что платформы могут развиваться быстрее, начиная с существующих

psychologically realistic computational modeling should become a core component of future public policy education.	исследовательских программ, чем путем разработки таких прогностических симуляций. ab initio. Все вместе эти преимущества имеют большое значение для того, чтобы поставить информативное моделирование в пределах досягаемости, так что, усиливая Сун (2018b) на этом этапе, психологически реалистичное вычислительное моделирование должно стать ключевым компонентом будущего образования в области государственной политики.
All of this aside, the most important upshot of the present proposal is that, at every level of analysis, there should be an accounting of energetic requirements and return into the ecology. This accounting should translate stress as understood on Tani's framework into social dynamics as represented in Sun's and Pereira's models, and we may compare the efficiencies of social arrangements operating under different types and rates of stress. In short, as stressors mount, performance becomes erratic. Some stressors even result in pathologies (as in Yamashita and Tani 2012, for example). Systems may run less efficiently or break down altogether. Policies must be adjusted, yet change must be accommodated. There are costs every step of the way. Modeling at select levels may inform policy makers on how to proactively mitigate these costs over near and long terms in such a way that implications may be made explicit, affording an opportunity for anticipatory response. For instance, impacts on the environment of a given policy initiative may be characterized at near and long terms, with more rapidly realized benefits evaluated against stress on affected populations, such that actions may be adjusted in coordination with other aspects of the larger system in terms of which the given policy is embedded. Simulations may confirm for instance that as industry slows, pollution declines, yet that standards of living eventually rise as, through a balance of high technology and traditional methods, people become wiser, healthier, live longer with greater leisure and with more time to reflect on the beauties of flourishing natural systems such as their own, and, without the chemical pollution and radiation threatening the natural world as is the case today, they flourish as well. However, moves must be made in the interim to pave the way to such a possible future. Simulations such as those subject of the present proposal may afford a survey of the landscape ahead, such that transitions through difficult passages may be most assured in getting there.	Помимо всего этого, наиболее важным результатом настоящего предложения является то, что на каждом уровне анализа необходимо вести учет энергетических потребностей и возвращаться в экологию. Этот учет должен переводить стресс, как он понимается в рамках Тани, в социальную динамику, представленную в моделях Сун и Перейры, и мы можем сравнивать эффективность социальных механизмов, работающих при различных типах и уровнях стресса. Короче говоря, по мере нарастания стрессора производительность становится неустойчивой. Некоторые стрессоры даже приводят к патологиям (например, в Ямашита и Тани 2012). Системы могут работать менее эффективно или вообще ломаться. Политика должна быть скорректирована, но изменения должны быть учтены. Есть расходы на каждом этапе пути. Моделирование на отдельных уровнях может информировать лиц, определяющих политику, о том, как заблаговременно смягчить эти затраты в краткосрочной и долгосрочной перспективе таким образом, чтобы последствия могли быть явными, предоставляя возможность для опережающего реагирования. Например, воздействие на окружающую среду данной политической инициативы может быть охарактеризовано в краткосрочной и долгосрочной перспективе, при этом более быстро реализуемые выгоды оцениваются по сравнению со стрессом для затронутого населения, так что действия могут корректироваться в координации с другими аспектами более крупной системы с точки зрения из которых данная политика встроена. Например, моделирование может подтвердить, что по мере замедления промышленности загрязнение снижается, но уровень жизни в конечном итоге повышается, так как благодаря балансу высоких технологий и традиционных методов люди становятся мудрее, здоровее, живут дольше с большим количеством свободного времени и имеют больше времени для размышлений о прелести процветающих природных систем, таких как их собственные, и без химического загрязнения и

	радиации, угрожающих природному миру, как это имеет место сегодня, они также процветают. Тем не менее, необходимо сделать шаги, чтобы проложить путь к такому возможному будущему. Имитационные модели, подобные тем, о которых идет речь в настоящем предложении, могут позволить провести обзор ландшафта впереди, так что переходы через трудные проходы могут быть наиболее надежными для достижения этой цели.
6 Conclusion	6. Заключение
With the present proposal in mind, we may read Putin's introductory prediction another way. Instead of increased capacities for violence and coercion, the mastery of AI and robotics may present us with opportunities to stabilize rather than to destabilize relations between seemingly disparate interests. This interpretation makes sense. For one thing, it accords with what cognitive science tells us about the nature of cognition. As seen in the brief review of Friston et al.'s research, and as confirmed in Tani et al.'s neurorobots, the human brain should not optimize to the existential uncertainty that results from a "race to the precipice" with wheels greased by headline garnering intelligent machine-mediated warfare (cf. Armstrong et al. 2016). Such a condition represents a diseased state, stressed to the breaking point. Rather, healthy cognition involves the minimization of this uncertainty. In this, we may find an indication as to what should be done.	Имея в виду настоящее предложение, мы можем прочитать вступительное предсказание Путина по-другому. Вместо повышения способности к насилию и принуждению, овладение искусственным интеллектом и робототехникой может дать нам возможность стабилизировать, а не дестабилизировать отношения между, казалось бы, несопоставимыми интересами. Такая интерпретация имеет смысл. Во-первых, это согласуется с тем, что когнитивная наука говорит нам о природе познания. Как видно из краткого обзора Фристона и соавт.это исследование, и как это было подтверждено в Тани и соавт. нейророботы, человеческий мозг не должен приспосабливаться к экзистенциальной неопределенности, которая возникает в результате "гонки к пропасти"с колесами, смазанными заголовками, собирающими интеллектуальную машинно-опосредованную войну (ср. Армстронг и соавт. 2016). Такое состояние представляет собой болезненное состояние, напряженное до предела. Напротив, здоровое познание предполагает минимизацию этой неопределенности. В этом мы можем найти указание на то, что должно быть сделано.
The emphasis of the preceding paper has not been on an AI arms race as competition, but rather on AI as affording avenues for peaceful cooperation toward ideal ends over the generations of human life and action that may be required to get us there. AI, to borrow from Ramstead et al. (2018), may help us to coordinate large-scale and long-term "phasetransitions" from unsustainable to sustainable, from unjust to just institutional arrangements proactively, openly, and in a non-coercive manner. Through the mastery of artificial intelligence thus, reason may yet rule the world after all.	Акцент в предыдущем документе был сделан не на гонке вооружений ИИ как конкуренции, а скорее на ИИ как обеспечении путей мирного сотрудничества для достижения идеальных целей на протяжении поколений человеческой жизни и действий, которые могут потребоваться для достижения этой цели. ИИ, чтобы заимствовать у Ramstead и соавт. ((2018), может помочь нам скоординировать крупномасштабные и долгосрочные "фазовые переходы" от неустойчивых к устойчивым, от несправедливых к справедливым институциональным механизмам проактивно, открыто и без принуждения. Таким образом, благодаря искусственному интеллекту разум все же может править миром.
Acknowledgements The author thanks Luis Pereira, Ron Sun, and Jun Tani	Благодарности Автор благодарит Луиса Перейру, Рона Суна и Джун

for help in the preparation of this draft, as well as Karamjit Gill for his constant support. Thanks are also due to Lorenzo Magnani for advice on abduction, the anonymous reviewers of this journal and attendees of the October 18, 2018 Tech and Values colloquium at the University of Twente for insights and objections informing final revisions.	Тани за помощь в подготовке этого проекта, а также Карамжита Гилла за его постоянную поддержку. Также следует поблагодарить Лоренцо Маньяни за советы по похищению, анонимных рецензентов этого журнала и участников коллоквиума Tech and Values 18 октября 2018 года в Университете Твенте за понимание и возражения, дающие представление об окончательных изменениях.
Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creat iveco mmons .org/licen ses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.	Открытый доступ Эта статья распространяется на условиях международной лицензии Creative Commons Attribution 4.0 (http: // creat iveco mmons .org / licen ses / by / 4.0 /), которая разрешает неограниченное использование, распространение и воспроизведение. на любом носителе, при условии, что вы предоставили соответствующую оценку первоначальному автору (авторам) и источнику, предоставили ссылку на лицензию Creative Commons и указали, были ли внесены изменения.
References	Использованная литература
Aliseda A (2006) Abductive reasoning. Logical investigations into discoveryand explanation. Springer, Berlin Armstrong S, Bostrom N, Shulman C (2016) Racing to the precipice: a model of artificial intelligence development. AI Soc 31:2:201–206 Cockshott P, Renaud K (2016) Humans, robots and values. Technol Soc 45:19–28 Fitzgerald THB, Dolan RJ, Friston K, Fitzgerald THB, Dolan RJ (2015) Dopamine, reward learning, and active inference. Front Comput Neurosci 9:1–16 Friston K (2010) The free-energy principle: a unified brain theory? Nat Rev Neurosci 11:127–138 Friston K (2012) A free energy principle for biological systems. Entropy 14(12):2100–2121 Friston K (2013) Life as we know it. J R Soc Interface 10:86–97 Friston K, FitzGerald T, Rigoli F, Schwartenbeck P, Pezzulo ODJ G (2016) Active inference and learning. Neurosci Biobehav Rev 68:862–879 Gabbay DM, Woods JH (2005) A practical logic of cognitive systems: insight and trial. Elsevier, Amsterdam Gok SE, Sayan E (2012) A philosophical assessment of computational models of consciousness. Cogn Syst Res 17–18:49–62 Han TA (2013) “Intention recognition, commitments and their roles	

in the evolution of cooperation: from artificial intelligence techniques to evolutionary game theory models" SAPERE 9. Springer, Berlin
Han TA, Pereira LM (2013) State-of-the-art of intention recognition and its use in decision making. AI Commun 26:237–246
Han TA, Pereira LM (2018) Evolutionary machine ethics. In: Bendel O (ed) Handbuch Maschinenethik. Springer reference Geisteswissenschaften. Springer, Wiesbaden, pp 1–25
Han TA, Pereira LM, Santos FC (2011) Intention recognition promotes the emergence of cooperation. Adapt Behav 3:264–279
Han TA, Pereira LM, Santos FC (2012) Corpus-based intention recognition in cooperation dilemmas. Artif Life 18(4):365–383
Han TA, Pereira LM, Santos FC, Lenaerts T (2013) Good agreements make good friends. Sci Rep 3:2695. https ://www.natur e.com/artic les/srep0 2695. Accessed 18 Oct 2018
Han TA, Pereira LM, Santos FC, Lenaerts T (2015) Emergence of cooperation via intention recognition, commitment, and apology—a research summary. AI Commun 2:709–715
Han TA, Pereira LM, Lenaerts T (2017) Evolution of commitment and level of participation in public goods games. Auton Agents Multi-Agent Syst 31(3):561–583
Ito M, Tani J (2004) On-line imitative interaction with a humanoid robot using a mirror neuron model. Proc IEEE Int Conf Robot Autom 2:1071–1076
Magnani L (2017) The abductive structure of scientific creativity: an essay on the ecology of cognition. Springer, Switzerland
Marr D (1982) Vision. WH Freeman, New York
Martinez-Vaquero LA, Han TA, Pereira LM, Lenaerts T (2015) Apology and forgiveness evolve to resolve failures in cooperative agreements. Sci Rep 5:10639
Martinez-Vaquero LA, Han TA, Pereira LM, Lenaerts T (2017) When agreement-accepting free-riders are a necessary evil for the evolution of cooperation. Sci Rep. https ://doi.org/10.1038/s4159 8-017-02625 -z
Murata S, Arie H, Ogata T, Sugano S, Tani J (2014) Learning to generate proactive and reactive behavior using a dynamic neural network model with time-varying variance prediction mecha-

nism. Adv Robot 28:1189–1203
Murata S, Masuda W, Tomioka S, Ogata T, Sugano S (2017) Mixing actual and predicted sensory states based on uncertainty estimation for flexible and robust robot behavior. In: Lintas A, Rovetta S, Verschure P, Villa A (eds) Artificial neural networks and machine learning—ICANN 2017. ICANN 2017. Lecture notes in computer science, vol 10613. Springer, Cham, pp 11–18
Namikawa J, Nishimoto R, Tani J (2011) A neurodynamic account of spontaneous behaviour. PLoS Comput Biol. https : //doi. org/10.1371/journ al.pcbi.10022 21
Newell A, Simon H (1976) Computer science as empirical inquiry: symbols and search. Commun ACM 19:113–126
Nishimoto R, Tani J (2009) Development of hierarchical structures for actions and motor imagery: a constructivist view from synthetic neuro-robotics study. Psychol Res 73:545–558
Paavola S (2006) On the origin of ideas: an abductivist approach to discovery. Department of Philosophy, University of Helsinki, Helsinki
Pereira LM, Saptawijaya A (2015) Bridging two realms of machine ethics. In: White J, Searle R (eds) Rethinking machine ethics in the age of ubiquitous technology. IGI Global, Hershey
Pereira LM, Saptawijaya A (2016) Programming machine ethics. Springer SAPERE series 26. Springer, Berlin
Pereira LM, Saptawijaya A (2017) Counterfactuals, logic programming and agent morality. In: Urbaniak R, Payette G (eds) Applications of formal philosophy: the road less travelled. Springer logic, argumentation and reasoning series. Springer, Berlin
Pereira LM, Dell'Acqua P, Pinto AM, Lopes G (2013) Inspecting and preferring abductive models. In: Nakamatsu K, Jain LC (eds) The handbook on reasoning-based intelligent systems. World Scientific Publishers, Singapore, pp 243–274
Peters A, McEwen BS, Friston K (2017) Uncertainty and stress: why it causes diseases and how it is mastered by the brain. Prog Neurobiol 156:164–188
Piaget J (1954) The construction of reality in the child. Basic Books, New York
Ramstead M, Badcock P, Friston K (2018) Answering Schrödinger's

question: a free-energy formulation. Phys Life Rev. https ://doi.org/10.1016/j.plrev .2017.09.001
RT (2017) 'Whoever leads in AI will rule the world': Putin to Russian children on Knowledge Day. https : //www.rt.com/news/40173 1 -ai-rule-world -putin /. Accessed 12 Oct 2018
Saptawijaya A, Pereira LM (2013) Towards practical tabled abduction in logic programs. In: Correia L, Reis LP, Cascalho J (eds) Progress in artificial intelligence. EPIA 2013. Lecture notes in computer science, vol 8154. Springer, Berlin, pp 223–234
Saptawijaya A, Pereira LM (2018) From logic programming to machine ethics. In: Bendel O (ed) Handbuch Maschinenethik. Springer, Berlin
Schwartenbeck P, FitzGerald THB, Mathys C, Dolan R, Kronbichler M, Friston K (2015) Evidence for surprise minimization over value maximization in choice behavior. Sci Rep 5:1–14
Shipp S, Adams RA, Friston KJ (2013) Reflections on agranular architecture: predictive coding in the motor cortex. Trends Neurosci 36(12):706–716
Simon HA (1977) Models of discovery: and other topics in the methods of science. D Reidel Pub Co, Dordrecht
Sun R (2002) Duality of the mind. Lawrence Erlbaum Associates, Mahwah
Sun R (2009) Motivational representations within a computational cognitive architecture. Cogn Comput 1(1):91–103
Sun R (2012) Grounding social sciences in cognitive sciences. MIT Press, Cambridge
Sun R (2013) Autonomous generation of symbolic representations through subsymbolic activities. Philos Psychol 26(6):888–912. https ://doi.org/10.1080/09515 089.2012.71103 5
Sun R (2016) Anatomy of the mind: exploring psychological mechanisms and processes with the Clarion cognitive architecture. Oxford University Press, New York
Sun R (2017) Potential of full human–machine symbiosis through truly intelligent cognitive systems. AI Soc. https ://doi.org/10.1007/s0014 6-017-0775-7
Sun R (2018a) Intrinsic motivation for truly autonomous agents. In: Abbass H, Scholz J, Reid D (eds) Foundations of trusted auton-

omy. Springer, Berlin, pp 273–292
Sun R (2018b) Cognitive social simulation for policy making. Policy Insights Behav Brain Sci 5(2):240–246
Sun R, Naveh I (2004) Simulating organizational decision-making using a cognitively realistic agent model. J Artif Soc Soc Simul. http://jasss .soc.surre y.ac.uk/7/3/5.html. Accessed 15 Dec 2017
Sun R, Wilson N (2014) A model of personality should be a cognitive architecture itself. Cogn Syst Res 29:1–30
Sun R, Merrill E, Peterson T (2001) From implicit skills to explicit knowledge: a bottom–up model of skill learning. Cogn Sci 25(2):203–244
Sun R, Coward LA, Zenzen MJ (2005) On levels of cognitive modeling. Philos Psychol 18(5):613–637
Sun R, Wilson N, Lynch M (2016) Emotion: a unified mechanistic interpretation from a cognitive architecture. Cogn Comput 8:1:1–14
Tani J (1996) Model-based learning for mobile robot navigation from the dynamical systems perspective. IEEE Trans Syst Man Cybern Part B-Cybern 26(3):421–436
Tani J (2007) On the interactions between top-down anticipation and bottom-up regression. Front Neurorobot 1:1–10
Tani J (2016) Exploring robotic minds: actions, symbols, and consciousness as self-organizing dynamic phenomena. Oxford University Press, New York
Tani J, Nolfi S (1999) Learning to perceive the world as articulated: an approach for hierarchical learning in sensory-motor systems. In: Pfeifer R, Blumberg B, Meyer JA, Wilson SW (eds) In: Proceedings of 5th international conference on simulation of adaptive behavior. MIT Press, Massachusetts, pp 270–279
Tani J, White J (2017) From biological to synthetic neurorobotics approaches to understanding the structure essential to consciousness, part 2. APA Newsl Philos Comput 16(2):29–41
White J (2016) Simulation, self-extinction, and philosophy in the service of human civilization. AI Soc 31(2):171–190
White J, Tani J (2016) From biological to synthetic neurorobotics approaches to understanding the structure essential to consciousness, part 1. APA Newsl Philos Comput 16(1):13–23

White J, Tani J (2017) From biological to synthetic neurorobotics approaches to understanding the structure essential to conscious- ness, part 3. APA Newsl Philos Comput 17(1):11–22 Yamashita Y, Tani J (2012) Spontaneous prediction error generation in Schizophrenia. PLoS One 5(5):e37843. https : //doi.org/10.1371/ journ al.pone.00378 43	